United States Court of Appeals for the Second Circuit

EXHIBITS

77-7145

UNITED STATES COURT OF APPEALS

FOR THE SECOND CIRCUIT

No. 77-7145

B

THE GUARDIANS ASSOCIATION, et al.,

Plaintiffs-Appellees,

-against-

CIVIL SERVICE COMMISSION OF THE CITY OF NEW YORK, et al.,

Defendants-Appellants.

UNITED STATES COURT OF APPEALS
FILED
APR 26 1977
A. DANIEL FUSARO, CLERK
SECOND CIRCUIT

On Appeal From an Order of the
United States District Court
for the Southern District of New York

EXHIBITS OF PLAINTIFFS-APPELLEES

VOLUME I

CHRISTOPHER CROWLEY
SARA E. STEINBOCK
1 Chase Manhattan Plaza
New York, New York 10005

OSCAR GARCIA-RIVERA
KENNETH KIMERLING
ROBERT H. HERMANN
Puerto Rican Legal Defense
and Education Fund, Inc.
95 Madison Avenue
New York, New York

Attorneys for Plaintiffs-Appellees

DEBORAH M. GREENBERG
10 Columbus Circle
New York, New York 10019

Of Counsel

PLAINTIFFS' EXHIBITS

1. Rand Report -- See, Supplemental Appendix of Plaintiffs-Appellees at page 85sa

2. EEOC Guidelines

3. APA Standards

4. Barrett Affidavit

5. Plaintiffs' Document Demand in Guardians I

6. Department of Personnel Notice of Examination, September 24, 1969

7. Marked for identification -- not proferred

8. Marked for identification -- not proferred

9. Bernard Orr -- Department of Personnel notice indicating exam taken

10. Jamie Felix -- Department of Personnel notice indicating exams

11. John Daley -- Notice of result of exam from Department of Personnel

12. Chart prepared by the Equal Opportunity Unit of the NYCPD showing number and racial composition of NYCPD as of 4/3/75

13. Chart prepared by the Equal Opportunity Unit of the NYCPD showing number and racial composition of NYCPD as of 8/28/75

14. Letter from Norman Redlich to Allan Luks of Rand Institute re proposed study

15. The New York City Rand Institute Proposal: Analysis Pertaining to Civil Action No. 72-928, 10/26/72

16. Proposal re Rand methodology -- Statistical Preparation for Guardians Association and Hispanic Society vs. The Civil Service Commission and the Police Department (Proposal)

17. Interim Report on Research on Police Trainee Exam #7046

18. Interim Report on Research on Police Trainee, Patrolman and Policewoman, Exam. Nos. 9080, 9081, April 13, 1971

19. Status of Study of exams Nos. 9080, 9081, (January 16, 1973)

20. Lavin Report

21. Test plan for exam 0013

22. Test plan for exam 9080

23. Test plan for exam 9049

24. Final test plan for exam 9080

25. Final test plan for exam 0013

26. Syllabus - police academy cucurriculum - July 1967

27, A-N Police academy exams

28. Marked for identification - not profferred

29. Marked for identification - not profferred

30. Guardians Association complaint to LEAA

31. Marked for identification - not profferred

32. September 8, 1975 letter from LEAA to Commissioner Codd

33. Guardians EEOC complaint

34. Marked for identification - not profferred

35. Costa Report, "Fair Employment Law and Minority Recruitment and Promotion within the New York City Police Department"

35 A. Costa Report (Plaintiffs Exhibit 35) with annotations by Deputy Commissioner Luis M. Neco

36. Memorandum from Peter Ring re Minority Recruitment, March 3, 1972

37. Memorandum from Charles Henry indicating rate of failures to return PA-15 form by race, December 15, 1971

38. Department of Personnel Statistice re scores received on 1973 exam by race

Ex 2

Equal Employment Opportunity Commission

Testing and Selecting Employees Guidelines

¶ 4010

The "Guidelines on Employment Testing Procedures," originally adopted by the Equal Employment Opportunity Commission on August 24, 1966, were superseded and enlarged upon by the "Guidelines on Employment Selection Procedures," which were codified on July 21, 1970 (29 CFR 1607), effective upon publication in the Federal Register on August 1, 1970 (35 F. R. 12333). Title 29—Labor, Chapter XIV, Part 1607, Sections 1607.1 to 1607.14.*

Table of Contents

Section		Paragraph
1607.1	Statement of Purpose	4010.1
1607.2	"Test" Defined	4010.2
1607.3	Discrimination Defined	4010.3
1607.4	Evidence of Validity	4010.4
1607.5	Minimum Standards for Validation	4010.5
1607.6	Presentation of Validity Evidence	4010.6
1607.7	Use of Other Validity Studies	4010.7
1607.8	Assumption of Validity	4010.8
1607.9	Continued Use of Tests	4010.9
1607.10	Employment Agencies and Employment Services	4010.10
1607.11	Disparate Treatment	4010.11
1607.12	Retesting	4010.12
1607.13	Other Selection Techniques	4010.13
1607.14	Affirmative Action	4010.14

[¶ 4010.1]

Section 1607.1. Statement of Purpose.—(a) The guidelines in this part are based on the belief that properly validated and standardized employee selection procedures can significantly contribute to the implementation of nondiscriminatory personnel policies, as required by Title VII. It is also recognized that professionally developed tests, when used in conjunction with other tools of personnel assessment and complemented by sound programs of job design, may significantly aid in the development and maintenance of an efficient work force and, indeed, aid in the utilization and conservation of human resources generally.

(b) An examination of charges of discrimination filed with the Commission and an evaluation of the results of the Commission's compliance activities has revealed a decided increase in total test usage and a marked increase in doubtful testing practices which, based on our experience, tend to have discriminatory effects. In many cases, persons have come to rely almost exclusively on tests as the basis for making the decision to hire, transfer, promote, grant membership, train, refer or retain, with the result that candidates are selected or rejected on the basis of a single test score. Where tests are so used, minority candidates frequently experience disproportionately high rates of rejection by failing to attain score levels that have been established as minimum standards for qualification.

It has also become clear that in many instances persons are using tests as the basis for employment decisions without evidence that they are valid predictors of employee job performance. Where evidence in support of presumed relationships between test performance and job behavior is lacking, the possibility of discrimination in the application of test results must be recognized. A test lacking demonstrated validity (i. e., having no known significant relationship to job behavior) and yielding lower scores for classes protected by Title VII may result in the rejection of many who have necessary qualifications for successful work performance.

(c) The guidelines in this part are designed to serve as a workable set of standards for employers, unions and employment agencies in determining whether their selection procedures conform with the obligations

* **Authority:** The provisions of this Part 1607 issued under Sec. 713, 78 Stat. 265, 42 U. S. C. § 2000e-12.

contained in Title VII of the Civil Rights Act of 1964. Section 703 of Title VII places an affirmative obligation upon employers, labor unions, and employment agencies, as defined in section 701 of the Act, not to discriminate because of race, color, religion, sex, or national origin. Subsection (h) of section 703 allows such persons "* * * to give and to act upon the results of any professionally developed ability test provided that such test, its administration or action upon the results is not designed, intended or used to discriminate because of race, color, religion, sex or national origin."

[¶ 4010.2]

Sec. 1607.2. "Test" Defined.—For the purpose of the guidelines in this part, the term "test" is defined as any paper-and-pencil or performance measure used as a basis for any employment decision. The guidelines in this part apply, for example, to ability tests which are designed to measure eligibility for hire, transfer, promotion, membership, training, referral or retention. This definition includes, but is not restricted to, measures of general intelligence, mental ability and learning ability; specific intellectual abilities; mechanical, clerical and other aptitudes; dexterity and coordination; knowledge and proficiency; occupational and other interests; and attitudes, personality or temperament. The term "test" includes all formal, scored, quantified or standardized techniques of assessing job suitability including, in addition to the above, specific qualifying or disqualifying personal history or background requirements, specific educational or work history requirements, scored interviews, biographical information blanks, interviewers' rating scales, scored application forms, etc.

[¶ 4010.3]

Sec. 1607.3. Discrimination Defined.—The use of any test which adversely affects hiring, promotion, transfer or any other employment or membership opportunity of classes protected by Title VII constitutes discrimination unless: (a) the test has been validated and evidences a high degree of utility as hereinafter described, and (b) the person giving or acting upon the results of the particular test can demonstrate that alternative suitable hiring, transfer or promotion procedures are unavailable for his use.

[¶ 4010.4]

Sec. 1607.4. Evidence of Validity.—(a) Each person using tests to select from among candidates for a position or for membership shall have available for inspection evidence that the tests are being used in a manner which does not violate § 1607.3. Such evidence shall be examined for indications of possible discrimination, such as instances of higher rejection rates for minority candidates than nonminority candidates. Furthermore, where technically feasible, a test should be validated for each minority group with which it is used; that is, any differential rejection rates that may exist, based on a test, must be relevant to performance on the jobs in question.

(b) The term "technically feasible" as used in these guidelines means having or obtaining a sufficient number of minority individuals to achieve findings of statistical and practical significance, the opportunity to obtain unbiased job performance criteria, etc. It is the responsibility of the person claiming absence of technical feasibility to positively demonstrate evidence of this absence.

(c) Evidence of a test's validity should consist of empirical data demonstrating that the test is predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated.

(1) If job progression structures and seniority provisions are so established that new employees will probably, within a reasonable period of time and in a great majority of cases, progress to a higher level, it may be considered that candidates are being evaluated for jobs at that higher level. However, where job progression is not so nearly automatic, or the time span is such that higher level jobs or employees' potential may be expected to change in significant ways, it shall be considered that candidates are being evaluated for a job at or near the entry level. This point is made to underscore the principle that attainment of or performance at a higher level job is a relevant criterion in validating employment tests only when there is a high probability that persons employed will in fact attain that higher level job within a reasonable period of time.

(2) Where a test is to be used in different units of a multiunit organization and no significant differences exist between units, jobs, and applicant populations, evidence obtained in one unit may suffice for the others. Similarly, where the validation process requires the collection of data throughout a multiunit organization, evidence of validity specific to each unit may not be required. There may also be in-

© 1972, Commerce Clearing House, Inc.

stances where evidence of validity is appropriately obtained from more than one company in the same industry. Both in this instance and in the use of data collected throughout a multiunit organization, evidence of validity specific to each unit may not be required: *Provided*, That no significant differences exist between units, jobs, and applicant populations.

[¶ 4010.5]

Sec. 1607.5. Minimum Standards for Validation.—(a) For the purpose of satisfying the requirements of this part, empirical evidence in support of a test's validity must be based on studies employing generally accepted procedures for determining criterion-related validity, such as those described in "Standards for Educational and Psychological Tests and Manuals" published by American Psychological Association, 1200 17th Street NW., Washington, D.C. 20036. Evidence of content or construct validity, as defined in that publication, may also be appropriate where criterion-related validity is not feasible. However, evidence for content or construct validity should be accompanied by sufficient information from job analyses to demonstrate the relevance of the content (in the case of job knowledge or proficiency tests) or the construct (in the case of trait measures). Evidence of content validity alone may be acceptable for well-developed tests that consist of suitable samples of the essential knowledge, skills or behaviors composing the job in question. The types of knowledge, skills or behaviors contemplated here do not include those which can be acquired in a brief orientation to the job.

(b) Although any appropriate validation strategy may be used to develop such empirical evidence, the following minimum standards, as applicable, must be met in the research approach and in the presentation of results which constitute evidence of validity.

(1) Where a validity study is conducted in which tests are administered to applicants, with criterion data collected later, the sample of subjects must be representative of the normal or typical candidate group for the job or jobs in question. This further assumes that the applicant sample is representative of the minority population available for the job or jobs in question in the local labor market. Where a validity study is conducted in which tests are administered to present employees, the sample must be representative of the minority groups currently included in the applicant population. If it is not technically feasible to include minority employees in validation studies conducted on the present work force, the conduct of a validation study without minority candidates does not relieve any person of his subsequent obligation for validation when inclusion of minority candidates becomes technically feasible.

(2) Tests must be administered and scored under controlled and standardized conditions, with proper safeguards to protect the security of test scores and to insure that scores do not enter into any judgments of employee adequacy that are to be used as criterion measures. Copies of tests and test manuals, including instructions for administration, scoring, and interpretation of test results, that are privately developed and/or are not available through normal commercial channels must be included as a part of the validation evidence.

(3) The work behaviors or other criteria of employee adequacy which the test is intended to predict or identify must be fully described; and, additionally, in the case of rating techniques, the appraisal form(s) and instructions to the rater(s) must be included as a part of the validation evidence. Such criteria may include measures other than actual work proficiency, such as training time, supervisory ratings, regularity of attendance and tenure. Whatever criteria are used they must represent major or critical work behaviors as revealed by careful job analyses.

(4) In view of the possibility of bias inherent in subjective evaluations, supervisory rating techniques should be carefully developed, and the ratings should be closely examined for evidence of bias. In addition, minorities might obtain unfairly low performance criterion scores for reasons other than supervisors' prejudice, as, when, as new employees, they have had less opportunity to learn job skills. The general point is that all criteria need to be examined to insure freedom from factors which would unfairly depress the scores of minority groups.

(5) Differential validity. Data must be generated and results separately reported for minority and nonminority groups wherever technically feasible. Where a minority group is sufficiently large to constitute an identifiable factor in the local labor market, but validation data have not been developed and presented separately for that group, evidence of satisfactory validity based on

other groups will be regarded as only provisional compliance with these guidelines pending separate validation of the test for the minority group in question. (See § 1607.9). A test which is differentially valid may be used in groups for which it is valid but not for those in which it is not valid. In this regard, where a test is valid for two groups but one group characteristically obtains higher test scores than the other without a corresponding difference in job performance, cutoff scores must be set so as to predict the same probability of job success in both groups.

(c) In assessing the utility of a test the following considerations will be applicable:

(1) The relationship between the test and at least one relevant criterion must be statistically significant. This ordinarily means that the relationship should be sufficiently high as to have a probability of no more than 1 to 20 to have occurred by chance. However, the use of a single test as the sole selection device will be scrutinized closely when that test is valid against only one component of job performance.

(2) In addition to statistical significance, the relationship between the test and criterion should have practical significance. The magnitude of the relationship needed for practical significance or usefulness is affected by several factors, including:

(i) The larger the proportion of applicants who are hired for or placed on the job, the higher the relationship needs to be in order to be practically useful. Conversely, a relatively low relationship may prove useful when proportionately few job vacancies are available;

(ii) The larger the proportion of applicants who become satisfactory employees when not selected on the basis of the test, the higher the relationship needs to be between the test and a criterion of job success for the test to be practically useful. Conversely, a relatively low relationship may prove useful when proportionately few applicants turn out to be satisfactory;

(iii) The smaller the economic and human risks involved in hiring an unqualified applicant relative to the risks entailed in rejecting a qualified applicant, the greater the relationship needs to be in order to be practically useful. Conversely, a relatively low relationship may prove useful when the former risks are relatively high.

[¶ 4010.6]

Sec. 1607.6. **Presentation of Validity Evidence.**—The presentation of the results of a validation study must include graphical and statistical representations of the relationships between the test and the criteria, permitting judgments of the test's utility in making predictions of future work behavior. (See § 1607.5(c) concerning assessing utility of a test.) Average scores for all tests and criteria must be reported for all relevant subgroups, including minority and nonminority groups where differential validation is required. Whenever statistical adjustments are made in validity results for less than perfect reliability or for restriction of score range in the test or the criterion, or both, the supporting evidence from the validation study must be presented in detail. Furthermore, for each test that is to be established or continued as an operational employee selection instrument, as a result of the validation study, the minimum acceptable cutoff (passing) score on the test must be reported. It is expected that each operational cutoff score will be reasonable and consistent with normal expectations of proficiency within the work force or group on which the study was conducted.

[¶ 4010.7]

Sec. 1607.7. **Use of Other Validity Studies.**—In cases where the validity of a test cannot be determined pursuant to § 1607.4 and § 1607.5 (e.g., the number of subjects is less than that required for a technically adequate validation study, or an appropriate criterion measure cannot be developed), evidence from validity studies conducted in other organizations, such as that reported in test manuals and professional literature, may be considered acceptable when: (a) The studies pertain to jobs which are comparable (i.e., have basically the same task elements), and (b) there are no major differences in contextual variables or sample composition which are likely to significantly affect validity. Any person citing evidence from other validity studies as evidence of test validity for his own jobs must substantiate in detail job comparability and must demonstrate the absence of contextual or sample differences cited in paragraphs (a) and (b) of this section.

[¶ 4010.8]

Sec. 1607.8. **Assumption of Validity.**—(a) Under no circumstances will the general reputation of a test, its author or its publisher, or casual reports of test utility be accepted in lieu of evidence of validity. Specifically ruled out are: assumptions of validity based on test names or descriptive labels; all forms of promotional literature;

© 1972, Commerce Clearing House, Inc.

data bearing on the frequency of a test's usage; testimonial statements of sellers, users, or consultants; and other nonempirical or anecdotal accounts of testing practices or testing outcomes.

(b) Although professional supervision of testing activities may help greatly to insure technically sound and nondiscriminatory test usage, such involvement alone shall not be regarded as constituting satisfactory evidence of test validity.

[¶ 4010.9]

Sec. 1607.9. Continued Use of Tests.—Under certain conditions, a person may be permitted to continue the use of a test which is not at the moment fully supported by the required evidence of validity. If, for example, determination of criterion-related validity in a specific setting is practicable and required but not yet obtained, the use of the test may continue: *Provided:* (a) The person can cite substantial evidence of validity as described in § 1607.7(a) and (b); and (b) he has in progress validation procedures which are designed to produce, within a reasonable time, the additional data required. It is expected also that the person may have to alter or suspend test cutoff scores so that score ranges broad enough to permit the identification of criterion-related validity will be obtained.

[¶ 4010.10]

Sec. 1607.10. Employment Agencies and Employment Services.—(a) An employment service, including private employment agencies, State employment agencies, and the U. S. Training and Employment Service, as defined in section 701(c), shall not make applicant or employee appraisals or referrals based on the results obtained from any psychological test or other selection standard not validated in accordance with these guidelines.

(b) An employment agency or service which is requested by an employer or union to devise a testing program is required to follow the standards for test validation as set forth in these guidelines. An employment service is not relieved of its obligation herein because the test user did not request such validation or has requested the use of some lesser standard than is provided in these guidelines.

(c) Where an employment agency or service is requested only to administer a testing program which has been elsewhere devised, the employment agency or service shall request evidence of validation, as described in the guidelines in this part, before it administers the testing program and/or makes referral pursuant to the test results. The employment agency must furnish on request such evidence of validation. An employment agency or service will be expected to refuse to administer a test where the employer or union does not supply satisfactory evidence of validation. Reliance by the test user on the reputation of the test, its author, or the name of the test shall not be deemed sufficient evidence of validity (see § 1607.8(a)). An employment agency or service may administer a testing program where the evidence of validity comports with the standards provided in § 1607.7.

[¶ 4010.11]

Sec. 1607.11. Disparate Treatment.—The principle of disparate or unequal treatment must be distinguished from the concepts of test validation. A test or other employee selection standard—even though validated against job performance in accordance with the guidelines in this part—cannot be imposed upon any individual or class protected by Title VII where other employees, applicants or members have not been subjected to that standard. Disparate treatment, for example, occurs where members of a minority or sex group have been denied the same employment, promotion, transfer or membership opportunities as have been made available to other employees or applicants. Those employees or applicants who have beeen denied equal treatment, because of prior discriminatory practices or policies, must at least be afforded the same opportunities as had existed for other employees or applicants during the period of discrimination. Thus, no new test or other employee selection standard can be imposed upon a class of individuals protected by Title VII who, but for prior discrimination, would have been granted the opportunity to qualify under less stringent selection standards previously in force.

[¶ 4010.12]

Sec. 1607.12. Retesting.—Employers, unions, and employment agencies should provide an opportunity for retesting and reconsideration to earlier "failure" candidates who have availed themselves of more training or experience. In particular, if any applicant or employee during the course of an interview or other employment procedure claims more education or experience, that individual should be retested.

[¶ 4010.13]

Sec. 1607.13. Other Selection Techniques.—Selection techniques other than tests, as defined in § 1607.2, may be improperly used so as to have the effect of discriminating against minority groups. Such techniques include, but are not restricted to, unscored or casual interviews and unscored application forms. Where there are data suggesting employment discrimination, the person may be called upon to present evidence concerning the validity of his unscored procedures as well as of any tests which may be used, the evidence of validity being of the same types referred to in §§ 1607.4 and 1607.5. Data suggesting the possibility of discrimination exist, for example, when there are differential rates of applicant rejection from various minority and nonminority or sex groups for the same job or group of jobs or when there are disproportionate representations of minority and nonminority or sex groups among present employees in different types of jobs. If the person is unable or unwilling to perform such validation studies, he has the option of adjusting employment procedures so as to eliminate the conditions suggestive of employment discrimination.

[¶ 4010.14]

Sec. 1607.14. Affirmative Action.—Nothing in these guidelines shall be interpreted as diminishing a person's obligation under both Title VII and Executive Order 11246 as amended by Executive Order 11375 to undertake affirmative action to ensure that applicants or employees are treated without regard to race, color, religion, sex, or national origin. Specifically, the use of tests which have been validated pursuant to these guidelines does not relieve employers, unions or employment agencies of their obligations to take positive action in affording employment and training to members of classes protected by Title VII.

The guidelines in this part are effective upon publication in the FEDERAL REGISTER [35 F. R. 12333, August 1, 1970].

© 1972, Commerce Clearing House, Inc.

standards

FOR EDUCATIONAL & PSYCHOLOGICAL TESTS

USA 33s - 475
(ED. 4-23-71)

PLAINTIFF

EXHIBIT
U. S. DIST. COURT
S. D. OF N. Y.

standards
FOR EDUCATIONAL & PSYCHOLOGICAL TESTS

Prepared by a joint committee of the
American Psychological Association
American Educational Research Association
National Council on Measurement in Education
Frederick B. Davis, *Chair*

Published by the
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1200 Seventeenth Street, N.W. • Washington, D.C. 20036

This book is a revision of the 1966 Standards for *Educational and Psychological Tests and Manuals*, published by the American Psychological Association.

Published by the American Psychological Association, Inc.
1200 Seventeenth Street, N.W., Washington, D.C. 20036

Office of Scientific Affairs
Miriam F. Kelty, *Administrative Officer*
Willo P. White, *Administrative Associate*

Office of Communications
Harold P. Van Cott, *Director*
Elliot R. Siegel, *Executive Editor for Separate Publications*
Pamela A. Taylor, *Editor*
Elizabeth V. Forrester, *Editor*

Copyright © 1974 by the
American Psychological Association, Inc.
1200 Seventeenth Street, N.W.
Washington, D.C. 20036

Library of Congress Catalog Card Number 74-75734

Printed in the United States of America

ds for *Educational and Psychological*
an Psychological Association.

Association, Inc.
n, D 20036

e Publications

4-75 4

FOREWORD

This document is the lineal descendant of *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, prepared by a committee of the American Psychological Association and published by that organization in 1954, and of *Technical Recommendations for Achievement Tests*, prepared by committees of the American Educational Research Association and the National Council on Measurements Used in Education and published by the National Education Association in 1955. In 1966, the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education published *Standards for Educational and Psychological Tests and Manuals* (through the American Psychological Association) to replace the two earlier documents.

By 1971, it had become apparent that problems and issues in testing, especially problems in selection for employment or for admission to educational institutions, had outdated the 1966 publication. Consequently, the Committee on Psychological Tests of the American Psychological Association (later merged with the Committee on Assessment) made contact with the American Educational Research Association and the National Council on Measurement in Education to form a joint committee for the revision of *Standards for Educational and Psychological Tests and Manuals*. The members of the joint committee were:

Philip Ash (APA)
Joan K. Bollenbacher (NCME)
Frederick B. Davis, Chairman (APA)
Robert L. Ebel (APA)
Edmund W. Gordon (APA)
Robert M. Guion (APA)
George F. Madaus (NCME)
William A. Mehrens (AERA)
Doris P. Mosby (APA)
James H. Ricks (AERA)
E. Belvin Williams (liaison with the APA Board of Scientific Affairs)

The principal author of *Standards for Educational and Psychological Tests* is Professor Robert M. Guion, Bowling Green State University, Bowling Green, Ohio. The committee expresses to him and to dozens of other psychologists and educators who provided comments and suggestions with respect to early drafts of the document its deep appreciation for their contributions. Special gratitude is owing to the following who generously read and commented on the final draft: Anne Anastasi, Nancy Cole, Lee J. Cronbach, Robert L. Linn, Julian C. Stanley, and Robert L. Thorndike. The joint committee is deeply indebted to Miriam F. Kelty and Willo P. White of the Central Office of the American Psychological Association for their assistance throughout the project.

Because some members of the joint committee judged that certain aspects of program evaluation and of the interpretation of scores derived from content-referenced tests have not been treated adequately within the scope of the present document, the committee has suggested that publication of a companion volume be considered by a succeeding joint committee.

Finally, the chairman of the committee wishes to express to members of the present committee his deep appreciation of their long hours of critical reading and committee discussion.

Frederick B. Davis
University of Pennsylvania

ittee wishes to express to members of the
n of their long hours of critical reading and

Frederick B. Davis
University of Pennsylvania

INTRODUCTION

In March, 1954, the American Psychological Association issued the *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, endorsed by the American Educational Research Association and the National Council on Measurement in Education. In January, 1955, the latter two organizations published a further document, *Technical Recommendations for Achievement Tests*. Subsequently, a joint committee of the three organizations consolidated, modified, and revised the two documents and in 1966, through the American Psychological Association, published the *Standards for Educational and Psychological Tests and Manuals*. The present document is both a revision and an extension of the 1966 *Standards*. It presents standards for test use as well as for test manuals; it is intended to guide both test developers and test users.

A test user is one who chooses tests, interprets scores, or makes decisions based on test scores. (People who do only routine administration or scoring of tests are not included in this definition, although test users often do both.) Test users include clinical or industrial psychologists, research directors, school psychologists, counselors, employment supervisors, teachers, and various administrators who select or interpret tests for their organizations. The audience for the *Standards* is, therefore, broad and cuts across publics with varying backgrounds and different training in measurement and statistics. Sections of the *Standards*, particularly those related to validity and reliability, are necessarily technical. These two sections should be meaningful to readers who have training approximately equivalent to a level between the master's degree and the doctorate in education or psychology. However, the remaining sections—the greater part of the document—are generally nontechnical and may be read with profit by all users.

The authors of the 1966 *Standards* declared that a test producer has an obligation to provide enough information about a test so that a qualified user will know what reliance can safely be placed on it; they also provided statements of consensus concerning the information that should be in a manual. It now appears desirable to provide similar statements of consensus concerning competency in testing practices.

Part of the stimulus for revision is an awakened concern about problems like invasion of privacy or discrimination against members of groups such as minorities or women. Serious misuses of tests include, for example, labeling Spanish-speaking children as mentally retarded on the basis of scores on tests standardized on "a representative sample of American children," or using a test with a major loading on verbal comprehension without appropriate validation in an attempt to screen out large numbers of blacks from manipulative jobs requiring minimal verbal communication.

These are specific examples of a general problem of test appro-

priateness. A test score describes but it does not explain a level of performance. Test performance may be influenced by many factors such as amount and quality of certain kinds of training, distractions during testing, sensory defects, inability to hear instructions because of poor administration, inappropriate language in instructions or in the test, inability to read, brain damage, motivation level, illumination level, cultural background of the examinee, or test-taking strategies.

Some unfairness may be built into a test, for example, requiring an inordinately high level of verbal ability to comprehend the instructions for a nonverbal test. Many of the social ills attributed to tests, however, seem more a result of the ways in which tests have been used than of characteristics of the tests themselves; for example, errors in administration, failure to consider the appropriateness of normative data, failure to choose an appropriate test, use of incorrect assumptions about the causes of a low or deviant test score, or administrative rigidity in using test scores for making decisions.

Tests and Test Uses to Which These Standards Apply

It is intended that these standards apply to any assessment procedure, assessment device, or assessment aid; that is, to any systematic basis for making inferences about characteristics of people.

A test is a special case of an assessment procedure. It may be thought of as a set of tasks or questions intended to elicit particular types of behavior when presented under standardized conditions and to yield scores that will have desirable psychometric properties such as high reliability and high validity.

Tests include standardized aptitude and achievement instruments, diagnostic and evaluative devices, interest inventories, personality inventories, projective instruments and related clinical techniques, and many kinds of personal history forms. It was pointed out in the 1966 *Standards* that the same general types of information are needed for all these varieties of published diagnostic, prognostic, and evaluative devices. It is equally appropriate to point out that unpublished assessment devices can be better used if the same kind of information is available to users.[1]

---

[1]It is sometimes suggested in response to perceptions of test abuse and unfair uses of tests that a moratorium on testing be observed until better and more appropriate instruments are developed and more equitable procedures can be instituted. The suggestion of such an extreme measure may be indicative of the growing sense of frustration and indignation felt particularly by some minority group members who sense that testing has had a disproportionately negative impact on their opportunities for equal access to success in education and employment. This suggestion, although well intended, seems futile for several reasons:

First, it fails to consider unfairness resulting from the misuses of tests. If new and better tests were subject to the same sorts of misuse, they might well produce the same sorts of errors (or errors of the same magnitudes) in the decisions based on them.

Second, it requires a corresponding but unlikely moratorium on decisions. Employers will continue to make employment decisions with or without standardized tests. Colleges and universities will still select students, some elementary pupils will still be recommended for special education, and boards of education will continue to evaluate the success of specific programs. If

behavior when presented under standardized conditions and to yield scores that will have desirable psychometric properties such as high reliability and high validity.

Tests include standardized aptitude and achievement instruments, diagnostic and evaluative devices, interest inventories, personality inventories, projective instruments and related clinical techniques, and many kinds of personal history forms. It was pointed out in the 1966 *Standards* that the same general types of information are needed for all these varieties of published diagnostic, prognostic, and evaluative devices. It is equally appropriate to point out that unpublished assessment devices can be better used if the same kind of information is available to users.[1]

[1] It is sometimes suggested in response to perceptions of test abuse and unfair uses of tests that a moratorium on testing be observed until better and more appropriate instruments are developed and more equitable procedures can be instituted. The suggestion of such an extreme measure may be indicative of the growing sense of frustration and indignation felt particularly by some minority group members who sense that testing has had a disproportionately negative impact on their opportunities for equal access to success in education and employment. This suggestion, although well intended, seems futile for several reasons:

First, it fails to consider unfairness resulting from the misuses of tests. If new and better tests were subject to the same sorts of misuse, they might well produce the same sorts of errors (or errors of the same magnitudes) in the decisions based on them.

Second, it requires a corresponding but unlikely moratorium on decisions. Employers will continue to make employment decisions with or without standardized tests. Colleges and universities will still select students, some elementary pupils will still be recommended for special education, and boards of education will continue to evaluate the success of specific programs. If those responsible for making decisions do not use standardized assessment techniques, they will use less dependable methods of assessment.

Third, tests are often useful for *finding* talent but are too often used only as devices for rejecting those with low scores; they can also be used to discover potential for performance that might not otherwise be observed. In this way, the use of tests may sometimes improve the prospects of minority group members and women.

There are wide variations in the sophistication of assessment techniques. At one extreme is the test that has gone through several revisions based on many research studies. Such a test may provide normative data based on thousands of cases classified into dozens of subpopulations. At the other extreme is the casual interview that provides assessments based on varying and unsystematically observed cues.

These standards are written specifically to apply to standardized tests. They apply in varying degrees, however, to the entire range of assessment techniques. If it is required that a relationship be demonstrated between scores (assessments) on an employment test and subsequent performance on a job, the requirement should in principle also apply to the judgments (assessments) of the employment interviewer. It may not be possible to apply the standards with the same rigor, but the kind of judgments the interviewer is to make can be identified; the time and procedures for developing and recording them can be standardized; and they can be validated in the same ways that scores are validated. When someone who makes personnel decisions developes his own assessment techniques (a practice not discouraged intentionally in these standards), he will find the standards useful guides for developing information similar to that in good test manuals; the principles are as relevant to him as to the professional test developer. If he chooses to use a test that has been developed by someone else, he may find the standards helpful in evaluating alternatives from which he may choose; moreover, the standards may help in developing a program of application.

There are many dimensions along which measuring instruments can be classified. Some are designed to measure abilities, some to measure accomplishments, others to measure attitudes or interests. Some are inventories, interview aids, biographical data forms, and experimental diagnostic devices, and are not called tests. Generally, however, the word "test" is used in these standards to apply to all kinds of measurement. What these different kinds have in common is that scores with desirable psychometric properties may be derived from each.

These standards also apply to *criterion measures*. Studies evaluating uses of well-developed tests too often employ inadequate criterion data. A criterion measure should have the psychometric properties expected of any other measurement, such as validity, including in special instances some form of criterion-related validity, for example, the relationship of an immediate criterion measure to an intermediate or more nearly ultimate measure. Criterion development should be guided by the standards guiding test development.

Some assessment techniques are used as interview aids. The intent of such use

is an idiographic analysis of an individual—an approach to assessment which places special reliance on the skill of the clinician. It is often argued that this use is so unlike the use of other testing procedures that it cannot be judged by essentially psychometric standards. The qualitative nature of the assessment is less the point at issue than the distinction that can be made between clinical and actuarial prediction. When tests, projective or otherwise, are used as aids to an interviewer's assessment, the interviewer is himself the final assessment device, and his assessments become the "scores." These assessments can and should be validated like other psychometric measures.

Component bits of information may be analyzed somewhat as items are analyzed. Proposals for arriving at idiographic interpretations are almost always based partially upon a nomothetic premise; for example, that a Rorschach determinant correlates with a specified internal factor. The usual standards can be applied to premises of this kind. Therefore, although interview aids can present unusual problems, their user requires the same information about them that he requires for any test score, and his use of them is subject to some of the same psychometric considerations (e.g., reliability) applicable to other test scores.

The developer of such an interview aid need not indicate his test's validity by correlating it with any simple criterion. But if he goes so far as to make any generalization about what "most people see" or what "schizophrenics rarely do," he is making a statistical claim and should be held to the usual rules for supporting it. Moreover, when on the basis of projective test data, biographical information, or various behavioral cues elicited during an interview the interviewer makes a statement such as "this man will fail or be subject to severe depression if placed in this situation," he is making a prediction based on his assessment and should be held to the standards for demonstrating the validity of his prediction.

A comment also seems appropriate about hidden tests (such as an interviewer's systematic attempt to assess a trait within the context of the interview) or other unobtrusive or observational measures. Some of these may raise ethical problems, but they do not differ, in principle, from other tests, and the standards apply as much to these unseen or unrecognized tests as to those more clearly perceived by the examinee. Therefore, the psychologist who counts examples of a specific type of response in a behavior-modification setting is as much responsible for the validity of his interpretations of change or the basic reliability of his observations as is any other test user.

In short, the standards are intended to be widely applicable both to standardized tests and a wide variety of other assessment techniques. The degree of applicability of individual standards to nontest assessments will vary; developers and users of such assessment procedures should at least observe the spirit of the standards.

Tests are used for basic research purposes as well as for practical purposes. Although these standards were not written with research purposes in mind, the

qualified
determine
ply to his
These
structional
ment; there
statements
procedures
analysis.

Informa
Guide

For each
manual, perh
provide enou
qualified user
regarding the
pretation of
quired prior to
test scores for

A manual is
its literal tru
pression it le
typical prof
manual is pr
an inaccurate
from the man
written. The
spirit and to
plemental pu
literal statem

A manual
information
Many tests
limited train
may not fol
understand
formation.
surement sp
tion on wh
adequacy
technical in

statistical claim and should be held to the usual rules for supporting it. Moreover, when on the basis of projective test data, biographical information, or various behavioral cues elicited during an interview the interviewer makes a statement such as "this man will fail or be subject to severe depression if placed in this situation," he is making a prediction based on his assessment and should be held to the standards for demonstrating the validity of his prediction.

A comment also seems appropriate about hidden tests (such as an interviewer's systematic attempt to assess a trait within the context of the interview) and other unobtrusive or observational measures. Some of these may raise ethical problems, but they do not differ, in principle, from other tests, and the standards apply as much to these unseen or unrecognized tests as to those more clearly perceived by the examinee. Therefore, the psychologist who counts examples of a specific type of response in a behavior-modification setting is as much responsible for the validity of his interpretations of change or the basic reliability of his observations as is any other test user.

In short, the standards are intended to be widely applicable both to standardized tests and a wide variety of other assessment techniques. The degree of applicability of individual standards to nontest assessments will vary; developers and users of such assessment procedures should at least observe the spirit of the standards.

Tests are used for basic research purposes as well as for practical purposes. Although these standards were not written with research purposes in mind, the qualified investigator should be able to determine the manner in which they apply to his research.

These standards cannot replace instructional material on test development; therefore, there are no specific statements directly related to such procedures as item writing or item analysis.

Information Standards as a Guide to Test Developers

For each test there should be a test manual, perhaps with supplements, to provide enough information for a qualified user to make sound judgments regarding the usefulness and interpretation of test scores. Research is required prior to the release of the test or test scores for operational use.

A manual is to be judged not merely by its literal truthfulness, but by the impression it leaves with the reader. If the typical professional user for whom the manual is prepared is likely to obtain an inaccurate impression of the test from the manual, the manual is poorly written. The standards apply to the spirit and tone of the manual (or supplemental publication) as well as to its literal statements.

A manual must often communicate information to many different groups. Many tests are used by people with limited training in testing. These users may not follow technical discussion or understand detailed statistical information. Other users are measurement specialists; they seek information on which to judge the technical adequacy of the test. Sometimes technical information can be presented in a supplementary handbook. Whatever the form, the prospective test user must have available to him the information needed for making whatever judgments his use of the test requires. Even when the test (or test battery) is developed for use within a single organization, a manual can often be helpful; preparation of a manual helps the test developer organize his thinking, codify his procedures, and communicate his ideas and intentions to his assistants.

It is not appropriate for this publication to call for a particular level of validity or reliability, or otherwise to establish technical test specifications for specific tests, but it is appropriate to ask that any test manual provide the information necessary for a test user to decide whether the consistency, relevance, or standardization of a test makes it suitable for his purpose. These standards need not prescribe minimum statistical specifications. Rather, their intent is to describe in an explicit and conveniently available form the information required by test users. In arriving at those requirements, it has been necessary to judge what is a reasonable compromise between pressures of cost and time, on the one hand, and the ideal, on the other. The test producer ordinarily spends large sums of money in developing and standardizing a test. Insofar as these recommendations indicate the kind of information that is most valuable to test users, authors and publishers can more efficiently allocate funds for gathering and reporting data of greatest value. Some provisions are more applicable than others in any specific case. The completion of predictive validity studies related to job criteria, for example, is

essential before a vocational interest inventory can be used properly, but it may be only desirable for a values inventory and irrelevant for an inventory designed to diagnose mental disorders. These standards, therefore, represent an attempt to state what type of studies should be completed before a test is ready for release to the profession. They can serve as a similar guide for those who are developing tests for their own use.

Procedural Standards as a Guide to Test Users

The test user, in selecting, administering, scoring, or interpreting a test, should know his purposes, what he is doing to achieve those purposes, and their probable consequences. It is not enough to have benign purposes; the user must know the procedures necessary to maximize effectiveness and to minimize unfairness in test use. He must evaluate the many factors that may have influenced test performance in light of his purposes. Where he finds that certain factors would unfairly influence performance, his procedures for using the test and interpreting the scores should be designed to minimize such influences.

Competence in test use is a combination of knowledge of psychometric principles, knowledge of the problem situation in which the testing is to be done, technical skill, and some wisdom. Although it is not appropriate to tell a test user that he needs particular levels of validity and reliability, it is appropriate to ask him to ascertain that his procedures result in reasonably valid predictions or reliable classifications, or otherwise conform to the purposes of his testing. These standards of practice are written more as guidelines than as commandments. It is as necessary to make cost-benefit compromises in test use as in test development. These standards provide useful guidelines for test users as well as for test developers.

Three Levels of Standards

Manuals can never give all the information, and test users can never follow all the procedures that might be desirable. At the same time, restricting this statement of recommendations solely to essential or indispensable information and practices might tend to discourage development and reporting of additional information. The standards are, therefore, grouped in three levels: Essential, Very Desirable, and Desirable. Each proposed requirement is judged based on its importance and the feasibility of attaining it.

The statements listed as Essential are intended to represent the consensus of present-day thinking concerning what is normally required for competent use of a test. Any test or testing situation may present some unique problems; it is undesirable for the standards to be treated as unduly rigid; for example, they should not bind the producer of a novel test to an inappropriate procedure or form of reporting. The Essential standards indicate what information or practices will be needed for most tests in most applications. When a test developer or test user fails to satisfy these requirements, he should do so only as a considered judgment. In any single test or testing

ions, or otherwise conform to the
ses of his testing. These stand-
of practice are written more as
lines than as commandments. It is
essary to make cost-benefit com-
ses test use as in test develop-
The standards provide useful
lines for test users as well as for
evelopers.

Three Levels of Standards

nuals can never give all the in-
atio and test users can never
w all the procedures that might be
able. At the same time, restricting
sta ment of recommendations
to ssential or indispensable in-
ation and practices might tend to
urage development and reporting
ddi nal information. The stan-
s a therefore, grouped in three
s: Essential, Very Desirable, and
rab Each proposed requirement
dge based on its importance and
easibility of attaining it.
he statements listed as Essential
nte ed to represent the consensus
pres t-day thinking concerning
t is normally required for com-
nt use of a test. Any test or testing
atio may present some unique
blem it is undesirable for the stan-
ls to be treated as unduly rigid; for
mple they should not bind the
duc of a novel test to an inap-
priate procedure or form of repor-
. The Essential standards indicate
t information or practices will be
ded or most tests in most ap-
ations. When a test developer or test
r fails to satisfy these requirements,
sho do so only as a considered
gme . In any single test or testing

situation, there may be some Essential standards that do not apply. It should be noted that many of these standards require thought rather than specific action as an outcome of thought; for example, "A test user should consider. . . ." In most cases, such statements are listed as Essential.

If some type of Essential information is not available on a given test, it is important to help the reader recognize that the research on the test is incomplete in this respect. A test manual should include clear statements of what research has been done and avoid misleading statements.

The category Very Desirable is used to draw attention to types of information or practices that contribute greatly to the user's understanding of the test and to competence in its use. Standards in this category have not been listed as Essential if their usefulness is debatable.

The category Desirable includes information and practices that are helpful but not Essential or Very Desirable.

When a test is widely used, the developer has a greater responsibility for investigating it thoroughly and providing more extensive reports about it than when the test is limited in use. Large sales make research financially possible. Therefore, the developer of a popular test can add information in subsequent editions of the manual. For tests having limited sales, it is unreasonable to expect that as much information will be furnished.

Cautions To Be Exercised in the Use of These Standards

Almost any test can be useful for some function and in some situations, but even the best test can have damaging consequences if used inappropriately. Therefore, the primary responsibility for the improvement of testing continues to rest on the shoulders of test users. It is hoped that these standards will be used to extend the professional training of many test users who are not now being trained appropriately. Professional training of personnel managers, school administrators, and classroom teachers should prepare them to better understand information about tests, test interpretations, and these standards. Such training will do much to improve the quality of test use and to minimize the extent of test misuse. The standards draw attention to recent developments in thinking about tests, test analysis, and test use. A comparison of these standards with those in earlier editions should remind test developers and test users that testing is a stable but not a static enterprise and that, in fact, there is room for improvement in the quality of assessments that are being made.

Tests are often developed and used in circumstances that lead to maintaining less than the highest standards of technical excellence. We do not intend to discourage those who must make assessments of people from doing the best they can with whatever training and collaborative resources are available to them. These standards, however, are written to promote excellence. They provide a kind of checklist of factors to be considered in designing, standardizing, validating, scoring, and interpreting tests. They may help test developers and test users decide what studies are needed and how those

studies might best be recorded in manuals or in validation reports. Test users who are not going to do independent research on a test should refer to these standards for guidance in the choice, administration, scoring, and interpretation of tests.

It is conceivable that a test developer could fulfill most of the standards presented and still produce a test that would fall short of his intended or stated objectives. Care should be exercised to adhere, both in test development and test use, to the spirit as well as to the letter of these standards. Because of the possibility of misunderstanding or misinterpretation, it would not be appropriate for test developers or test users to state that a manual or procedure "satisfies" or "follows" these standards. There would be no objection to a statement that one has "taken into account or considered" these standards.

A final caveat is necessary in view of the prominence of testing issues in litigation. This document is prepared as a technical guide for those within the sponsoring professions; it is *not* written as law. What is intended is a set of standards to be used in part for self-evaluation by test developers and test users. An evaluation of their competence does not rest on the literal satisfaction of every relevant provision of this document. The individual standards are statements of ideals or goals, some having priority over others. Instead, an evaluation of competence depends on the degree to which the intent of this document has been satisfied by the test developer or user.

s

ion to a statement that one has ken into account or considered" se standards.

A fin caveat is necessary in view of prominence of testing issues in gatio This document is prepared as echn al guide for those within the nsoring professions; it is *not* written aw. What is intended is a set of stand- s t be used in part for self- luat by test developers and test rs. An evaluation of their com- nce does not rest on the literal sfac n of every relevant provision his cument. The individual stand- s are statements of ideals or goals, e h ing priority over others. In- d, evaluation of competence ends on the degree to which the in- of this document has been satisfied he t developer or user.

STANDARDS FOR TESTS, MANUALS, AND REPORTS

Tests vary in the amount of knowledge and research required to develop them. Much background work is needed for a test that is published or otherwise distributed for widespread use. Less work need be done for a test developed for local use. At any level, however, better tests and testing can be expected where test developers have been guided by fundamental considerations and have demonstrated this in writing.

A test user needs information describing a test's rationale, development, technical characteristics, administration, and interpretation. Such information is ordinarily expected in a test manual or in its supplements. This information is also needed by those using a test or test battery that has not been published but which is used within an organization to aid in making decisions. For these uses, a properly prepared manual reports to local users and to other interested persons (colleagues in other organizations, representatives of governmental agencies, representatives of citizen's groups, etc.) the procedures followed in construction of the test, in its use, and in the interpretation of scores derived from it. In certification or selection programs, a manual can present information about the program as a whole as well as about component tests. Data supporting claims for the program, procedures followed, kinds of tests used, and related information should be recorded not only to provide an adequate basis for the proper use of tests but also to make the information available for public scrutiny.

The development of a test or testing program is based on research; the report of that research is often contained in a manual. These standards, therefore, concentrate on the manual (and any supplementary publication) as the full and proper report of what was done in test development; they specify standards of reporting from which one may infer standards for research.

A. Dissemination of Information

A test user needs information to help him use the test in standard ways and to evaluate a test relative to others he might select for a given purpose. The information that he needs to select a test or to use it must come, at least in part, from the test developer. Practices of authors and publishers in furnishing information have varied. Sometimes the test manual offers only vague directions for administering and scoring, norms of uncertain origin, and perhaps nothing more. In contrast, some manuals furnish extensive information on test development, validity, reliability, bases for normative information, appropriate kinds of interpretations and uses, and they present all such information in detail.

A1. When a test is published or otherwise made available for operational use, it should be accompanied by a manual (or other published or readily available

information) that makes every reasonable effort to follow the recommendations of these standards and, in particular, to provide the information required to substantiate any claims that have been made for its use. Essential

[Comment: The term "operational use" refers to making practical decisions about the evaluation or handling of individuals, groups, curricula, therapeutic treatments, and so on.

The term "manual" refers to documents describing procedures of test development, use, interpretation, relevant research, normative data, and related information. Depending on such things as the amount of information to report and the diversity of uses and users, the term may designate a document entirely within one cover or a series of separately bound pamphlets. This term might also be extended to include procedural manuals governing the use of tests or of test batteries in, for example, selection situations; the wording and importance of many of these standards would be different for a procedural manual, but the principles applicable to test manuals would at least, therefore, be analogous.

Not all of the standards in this report will apply to any one particular test. A standard may be ignored if it is irrelevant in light of the purpose of the test and the claims made for it, but it may not be ignored merely because it is difficult to meet or has not usually been met by a similar test.]

A1.1. If information needed to support interpretations suggested in the manual cannot be presented at the time the manual is published, the manual should satisfy the intent of standard A1 by pointing out the absence and importance of this information. Essential

A1.2. Where the information is too extensive to be fully reported in the manual, the essential information should be summarized and accompanied by references to other sources of information, such as technical supplements, articles, or books. Very Desirable

[Comment: Developers of some well-known tests provide extensive technical manuals, make further research data available through other sources (such as the Education Resources Information Center), prepare annotated bibliographies, or include relevant information in technical books which users are encouraged to consult. In other instances, the essential information is given in the manual sold with the instrument, along with references to other useful sources.

Publications by persons other than the author of the test frequently fulfill many functions of a manual. If a book about a test is designed to serve as a manual, its author and publisher have the same responsibility in preparing it as do the author and publisher of a test.]

A1.2.1. When information about a test is provided in a separate publication, that publication should meet the same standards of accuracy and freedom from misleading impressions that apply to the manual. Essential

A1.2.2. Promotional material for a

should satisfy the intent of standard A1 by pointing out the absence and importance of this information. Essential

A1.2. Where the information is too extensive to be fully reported in the manual, the essential information should be summarized and accompanied by references to other sources of information, such as technical supplements, articles, or books. Very Desirable

[Comment: Developers of some well-known tests provide extensive technical manuals, make further research data available through other sources (such as the Education Resources Information Center), prepare annotated bibliographies, or include relevant information in technical books which users are encouraged to consult. In other instances, the essential information is given in the manual sold with the instrument, along with references to other useful sources.

Publications by persons other than the author of the test frequently fulfill many functions of a manual. If a book about a test is designed to serve as a manual, its author and publisher have the same responsibility in preparing it as do the author and publisher of a test.]

A1.2.1. When information about a test is provided in a separate publication, that publication should meet the same standards of accuracy and freedom from misleading impressions that apply to the manual. Essential

A1.2.2. Promotional material for a test should be accurate and should not give the reader false impressions. Essential

[Comment: One publisher presents an extensive and complete bibliography, without comment or annotation, of research involving a test; he does *not* mention that many of the entries are studies with negative findings. The impression is one of extensive use, not of limitations to the usefulness of the test.]

A1.2.3. Informational material distributed within a using organization should be accurate, complete for the purposes of the reader's need, and written in language that will not give the reader a false impression. Essential

[Comment: Such information is often given in brief memoranda. In preparing these brief reports, the technical capability of the readers may be kept in mind, but this does not suggest that essential information be either omitted or distorted in the interest of simplicity. Where a reader may be expected to receive such reports regularly, efforts can be made to increase his ability to understand technical detail.]

A2. A test manual should describe fully the development of the test: the rationale, specifications followed in writing items or selecting observations, and procedures and results of item analysis or other research. Essential

A2.1. Data gathered during the process of developing a test before it is in final form should be clearly distinguished from data pertaining to the test in final form. Essential

A2.2. A test manual should specify the need for maintaining necessary test security. Very Desirable

[Comment: For example, a manual might describe some acceptable coaching practices. If so, it would be appropriate to add warnings against unacceptable practices that might jeopardize test security.]

A2.3. A test manual or supplementary document should provide representative sample items and a statement of the intended purpose of the test in a form that can be made available to those concerned about the nature and quality of a testing program. Very Desirable

[Comment: The evaluation of a test may not fall exclusively to those who are technically trained. Examinees, members of citizen panels, civil rights advocates, and parents are among those who may have reason to make judgments about the appropriateness of a test. Their right to do so need not conflict with the necessity to maintain test security if descriptive and explanatory materials are made available.

One publisher of educational tests has published descriptive material in nontechnical language for a wide variety of tests; pamphlets include information on test development and rationale as well as examples of items and suggestions on test-taking strategies.]

A2.4. The identity and professional qualifications of item writers and editors should be described in instances where they are relevant; for example, when adequacy of coverage of a subject-matter achievement test cannot appropriately or practically be measured

against any external criterion. Desirable

A3. The test and its manual should be revised at appropriate intervals. The time for revision has arrived whenever changing conditions of use or new research data make any statements in the manual incorrect or misleading. Very Desirable

[Comment: The technical characteristics and the appropriateness of a test may change as social conditions and attitudes, job definitions, educational pressures, or the composition of relevant school populations change.]

A3.1. Competent studies of the test following its publication, whether the results are favorable or unfavorable to the test, should be taken into account in revised editions of the manual or its supplementary reports. Pertinent studies by investigators other than the test authors and publishers should be included. Very Desirable

[Comment: The developer of one test has published a comprehensive review of validity studies of the test covering a 15-year period.]

A3.2. When the test is revised or a new form is issued, the manual should be suitably revised to take those changes into account. In addition, the nature and extent of the revision and the comparability of data from the old test and the revised test should be explicitly stated. Essential

[Comment: It is useful for publishers to identify revisions of test manuals in their catalogs and to take other steps to increase the probability that test users have current information.]

A3.2.1. If a short form of a test is prepared by reducing the number of items or organizing a portion of the test into a separate form, new evidence should be obtained and reported for that shorter test. Essential

[Comment: It is especially important to report the reliability and other technical data for the test in its shorter form, since placing items in a new context may alter responses to them.

In the manual for one test that has two alternate forms, the validity data presented were obtained using the sum of the scores of the two forms. It would have been more appropriate to have presented the data for each form independently.]

A3.2.2. When a short form is prepared from an established test, the manual should present evidence that the items in the short form represent the items in the long form or measure the same characteristics as the long form. Very Desirable

[Comment: When no short form of a test has been prepared but there is reason to believe that it is commonly used in a shortened form, the manual should remind the reader that data in the manual may not be applicable to results of administration of a shortened form.

One revision of a long-established achievement test battery illustrates a desirable practice by listing all previous editions and then describing in detail the relation of the new revision to the previous editions.]

increase the probability that test users have current information.]

A3.2.1. If a short form of a test is prepared by reducing the number of items or organizing a portion of the test into a separate form, new evidence should be obtained and reported for that shorter test. Essential

[Comment: It is especially important to report the reliability and other technical data for the test in its shorter form, since placing items in a new context may alter responses to them.

In the manual for one test that has two alternate forms, the validity data presented were obtained using the sum of the scores of the two forms. It would have been more appropriate to have presented the data for each form independently.]

A3.2.2. When a short form is prepared from an established test, the manual should present evidence that the items in the short form represent the items in the long form or measure the same characteristics as the long form. Very Desirable

[Comment: When no short form of a test has been prepared but there is reason to believe that it is commonly used in a shortened form, the manual should remind the reader that data in the manual may not be applicable to results of administration of a shortened form.

The revision of a long-established achievement test battery illustrates a desirable practice by listing all previous editions and then describing in detail the relation of the new revision to the previous editions.]

B. Aids to Interpretation

The responsibility for making inferences about the meaning and legitimate uses of test results rests primarily with the user. In making such judgments, however, he must depend in part on information about the test made available by its developer.

The manual or report form from a scoring service cannot fully prepare the user for interpreting the test. He will sometimes have to make judgments that have not been substantiated by published evidence. Thus, the vocational counselor cannot expect to have validity data available for each job about which he makes tentative predictions from test scores. The counselor or employment interviewer will have examinees who do not fit into any group for which normative or validity data are available. The teacher will have to evaluate the content of an achievement test in terms of his instructional goals and emphasis. The clinician must bring general data and theory into his interpretation of data from a personality inventory. The degree to which the manual can be expected to prepare the user for accurate interpretation and effective use of the test varies with the type of test and the purpose for which it is used. It is the test developer's responsibility to provide the information necessary for good judgment; in fact, developers should make tests as difficult to misuse and to misinterpret as they can.

B1. The test, the manual, the record forms, and other accompanying material should help users make correct interpretations of the test results and should warn against common misuses. Essential

B1.1. Names given to published tests, and to parts within tests, should be chosen to minimize the risk of misinterpretation by test purchasers and subjects. Essential

[Comment: It is desirable that names carry no unwarranted suggestion as to the characteristics measured. Such descriptions as "culture-free," "intelligence," "introversion," "creativity," "primary mental abilities," or "productivity quotients" are questionable for published tests, unless there is appropriate evidence of construct validity, since they may suggest interpretations going beyond the demonstrable meaning of the scores.]

B1.1.1. Devices for identifying interests and personality traits through self-report should be entitled "inventories," "questionnaires," or "checklists" rather than "tests." Very Desirable

[Comment: In referring to such instruments in textual material, however, as in these standards, the word "test" may be used to simplify the language even where it is properly avoided in the title.]

B1.2. The manual should draw the user's attention to data that especially need to be taken into account in the interpretation of test scores. Very Desirable

[Comment: Many test manuals point out variables that should be considered in the interpretation of a test score, such as information about school record, recommendations, or clinically relevant history.

A personality assessment manual may provide data to show that the psychologist should consider such facts as the sex and age of the subject, whether his parents are dead or separated, the ages and sexes of his siblings, or his vocational or marital status.]

B1.3. The manual should call attention to marked influences on test scores known to be associated with region, socioeconomic status, race, creed, color, national origin, or sex. Essential

[Comment: Social or cultural factors known to affect performance on the test differentially, administrator errors that are frequently repeated, examiner-examinee differences, and other factors that may result in spurious or unfair test scores should, for example, be clearly and prominently identified in the manual.]

B1.4. The manual should draw attention to, and warn against, any serious error of interpretation that is known to be frequent. Essential

[Comment: Some users of general intelligence tests think of the score as a direct measure of inherent native ability, given and unchanging; manuals of such tests may be expected to caution against this interpretation and to do so with reference to appropriate data. They should clearly warn users against unwarranted assumptions about the generality of normative data, particularly avoiding the impression that national norms are genuinely representative when in fact they are not.

Manuals for interest inventories can apply this standard by stressing the fact that interest does not necessarily imply ability and is only one of many factors to be considered in choosing among occupations.]

B2. The test manual should state explicitly the purposes and applications for which the test is recommended. Essential

[Comment: A clear statement of a test's purposes will help prevent the misapplication of test scores. It will alert the user to the kind and extent of evidence he should expect to find in the manual in support of the claims made for the test by the author and publisher. For example, if an achievement test is recommended as a survey test of what students know, an accurate description of its content is important. If, on the other hand, it is recommended as a diagnostic test or one that predicts performance, data on its relationship with one or more criteria are required. See Section E on validity.]

B2.1. If a test is intended for research use only and is not distributed for operational use, that fact should be prominently stated in the accompanying materials. Essential

[Comment: If the developer of a new device (e.g., for studying personality) releases his instrument for studies by other investigators before he considers it ready for operational use, it is appropriate to print "distributed for research use only" on the test package, on the cover of the booklet of directions, and in any catalog where it is listed. This cautions against premature use of the instrument in guidance or selection.]

B3. The test manual sh[illegible] clearly the psychological, [illegible] or other reasoning under[illegible] and nature of the chara[illegible] intended to measure. E[illegible]

[Comment: There [illegible] explicit reasons for setti[illegible] it has been done; it [illegible] that certain psychologi[illegible] required in taking the [illegible] certain traits are being [illegible] result. The identifica[illegible] processes may be bas[illegible] empirical research. [illegible] processes internal to th[illegible] any case, a clear des[illegible] construct or content and [illegible] of measurement enables [illegible] the test by its confor[illegible] psychological or educa[illegible] well as by statistical [illegible] efficacy.]

B3.1. In the case of [illegible] for content-referenced [illegible] special attention sho[illegible] defining the conte[illegible] operational terms. In [illegible] mastery test, the [illegible] rationale for any cutt[illegible] suggests should be [illegible] procedures that the u[illegible] to establish mastery [illegible] described. Essential

[Comment: The test [illegible] information so that be[illegible] concept of mastery or [illegible] that of the test author.]

B4. The test manual s[illegible] any special qualifica[illegible] administer the test and [illegible] properly. Essential

interest does not necessarily imply ity and is only one of many factors be nsidered in choosing among upa ns.]

32. The test manual should state licit the purposes and applications wh h the test is recommended. ential

Com ent: A clear statement of a 's rposes will help prevent the sapplication of test scores. It will rt t user to the kind and extent of den he should expect to find in the nual in support of the claims made the test by the author and publisher. r e nple, if an achievement test is com end d as a survey test of what dents know, an accurate description its content is important. If, on the her nd, it is recommended as a agn ic test or one that predicts rformance, data on its relationship th or more criteria are required. e S ion E on validity.]

B2.1. If a test is intended for sea use only and is not distributed r o ational use, that fact should be rom ently stated in the accompany- g materials. Essential

[C ment: If the developer of a new evic (e.g., for studying personality) eleases his instrument for studies by the nvestigators before he considers r y for operational use, it is ppropriate to print "distributed for esearch use only" on the test package, n t cover of the booklet of directions, nd any catalog where it is listed. This cautions against premature use of he instrument in guidance or sele on.]

B3. The test manual should describe clearly the psychological, educational, or other reasoning underlying the test and nature of the characteristic it is intended to measure. Essential

[Comment: There ordinarily are explicit reasons for setting up the test as it has been done; it may be assumed that certain psychological processes are required in taking the test and that certain traits are being measured as a result. The identification of these processes may be based on a theory, empirical research, or empirical processes internal to the test itself. In any case, a clear description of the construct or content and of the manner of measurement enables a user to judge the test by its conformity to his own psychological or educational insight as well as by statistical evidence of its efficacy.]

B3.1. In the case of tests developed for content-referenced interpretation, special attention should be given to defining the content domain in operational terms. In the case of a mastery test, the test developer's rationale for any cutting score that he suggests should be specified, or the procedures that the user might employ to establish mastery levels should be described. Essential

[Comment: The test user needs such information so that he can compare his concept of mastery or competence with that of the test author.]

B4. The test manual should identify any special qualifications required to administer the test and to interpret it properly. Essential

[Comment: One manual differentiates psychologists who work with children from those who work only with adults in identifying qualifications needed to use an individually administered test for children. Another offers specifications for administering the test to non-English-speaking students.

User qualifications might be described in terms of special training generally thought necessary to achieve competence. It may be possible for some test manuals to identify the most frequent sources of error in test use and to specify the kind of user training necessary to eliminate these common errors.

B4.1. The test manual should not imply that a test is "self-interpreting." It should specify information to be given about test results to persons who lack the training usually required to interpret them. Essential

[Comment: It is not ordinarily desirable to entrust interpretation of scores to an untrained person. There are, of course, tests that can be scored by the examinee, and it is often useful to give scores to students or parents. Where these practices are followed, the sense of this standard is that interpretative aids should also be given.

The manual should indicate what may be done by untrained persons and what should not be done. The manual for one well-known interests test, for example, indicates that examinees may perform the mechanics of scoring their own tests but properly stresses that they need the help of a trained teacher or counselor in making interpretations and future plans.]

B4.2. Where a test is recommended for a variety of purposes or types of inference, the manual should indicate the amount of training required for each use. Essential

B4.3. The manual should draw the user's attention to references with which he should become familiar before attempting to interpret the test results. Very Desirable

[Comment: The references might be to books or articles dealing with related psychological theory or with the particular test in question.]

B5. Evidence of validity and reliability, along with other relevant research data, should be presented in support of any claims being made. Essential

[Comment: Standards for validity and reliability are extensive. Moreover, they are as applicable to research reports prepared by test users as to test manuals. For these reasons, and because of their overreaching importance, a major section of this document presents standards for reports of research on validity and reliability. Adherence to the intent of this standard requires adherence to the appropriate standards in that section.]

B5.1. Statements in the manual reporting relationships are by implication quantitative and should be stated as precisely as the data permit. If data to support such statements have not been collected, that fact should be made clear. Essential

[Comment: Writers sometimes say, for example, "Spatial ability is required for architectural engineering," or, "Bizarre responses may indicate schizophrenic tendencies." Such statements by themselves are quantitatively inadequate. In what proportion of cases giving bizarre responses has schizophrenia been shown to develop? How much has architectural success been found to depend upon spatial ability? Numerical data relating the test scores to definite criteria would help to provide the answers.]

B5.2. Statistical procedures that are well known and readily interpreted should be preferred for reporting any quantitative information. Any uncommon statistical techniques should be explained, and references to descriptions of them should be given. Essential

[Comment: Publishers need not uniformly adhere to the procedures commonly used for reporting data, but terminology and procedures should be sufficiently common in practice to permit adequate judgment by reasonably competent users.

This standard is an elaboration of the principle that data presented in a manual should not be misleading. For example, it is misleading to show the value of combining tests in a battery in a regression equation by using data where intercorrelations are lower than those reported elsewhere in the manual.]

B5.3. When the statistical significance of a relationship is reported, the statistical report should be in a form that makes clear the sensitivity or power of the significance test. Essential

[Comment: Statistical significance

"bizarre responses may indicate schizophrenic tendencies." Such statements by themselves are quantitatively inadequate. In what proportion of cases giving bizarre responses has schizophrenia been shown to develop? How much has architectural success been found to depend upon spatial ability? Numerical data relating the test scores to definite criteria would help to provide the answers.]

B5.2. Statistical procedures that are well known and readily interpreted should be preferred for reporting any quantitative information. Any uncommon statistical techniques should be explained, and references to descriptions of them should be given. Essential

[Comment: Publishers need not uniformly adhere to the procedures commonly used for reporting data, but terminology and procedures should be sufficiently common in practice to permit adequate judgment by reasonably competent users.

This standard is an elaboration of the principle that data presented in a manual should not be misleading. For example, it is misleading to show the value of combining tests in a battery in a regression equation by using data where intercorrelations are lower than those reported elsewhere in the manual.]

B5.3. When the statistical significance of a relationship is reported, the statistical report should be in a form that makes clear the sensitivity or power of the significance test. Essential

[Comment: Statistical significance

that has no practical usefulness can often be obtained by using a very large number of cases. For example, a well-known inventory yields statistically significant differences between large samples of males and of females, but the differences are too small to be of practical importance. Conversely, one who uses an insensitive statistical test can falsely conclude that there is no difference of practical importance. In general, it is more appropriate in reporting test data to state a confidence interval or the likelihood function for the parameter of interest than to report only that the null hypothesis can or cannot be rejected.]

B5.4. The manual should differentiate between an interpretation that is applicable only to average tendencies of a group and one that is applicable to an individual within the group. Very Desirable

B5.5. The manual should state clearly what interpretations are intended for each subscore as well as for the total test. Essential

[Comment: Where subscores are obtained only for convenience in scoring the test, and no interpretation is intended, this should be made clear. For some tests, keys are provided for subscores that have possible research use but are not intended to be interpreted; this should be made clear.]

B6. Test developers or others offering computer services for test interpretation should provide a manual reporting the rationale and evidence in support of computer-based interpretations of scores. Essential

[Comment: A computer makes possible the storage and recall of large amounts of data; test interpretation can be greatly assisted by the use of computer data banks. Computer scoring services may provide lengthy printouts of descriptive and prognostic information from individual profiles on a test battery or personality inventory. The user of such printouts needs to know the reasoning and the evidence supporting the suggested interpretations because they are as fallible as other subjective interpretations.]

C. Directions for Administration and Scoring

Interpretations of test and measurement techniques, like those of experimental results, are most reliable when the measurements are obtained under standardized or controlled conditions. To be sure, there are circumstances in testing where it may be important to change conditions systematically for maximum understanding of the performance of an individual. For example, an examiner may systematically modify procedures in successive readministrations of a test to explore the limits of a child's mastery of a specific content area such as a set of concepts. Nevertheless, the test developer should provide a standard procedure from which modifications can be made. Without standardization, the quality of interpretations will be reduced to whatever extent differences in procedure influence performance.

For most purposes, great emphasis is properly placed on strict standardization of procedures for administering a test and reciting its instructions. If a test is to be used for a wide range of subpopulations, these pro-

cedures should be wholly comprehensible to all examinees in each subpopulation.

C1. The directions for administration should be presented in the test manual with sufficient clarity and emphasis so that the test user can duplicate, and will be encouraged to duplicate, the administrative conditions under which the norms and the data on reliability and validity were obtained. Essential

[Comment: Because persons administering tests in schools and industry sometimes may not follow instructions rigidly and may not understand the need for doing so, it is necessary that the manual be insistent and persuasive on this point. Some tests are fully administered by tape recordings to insure standardization of procedure.]

C1.1. The directions published in the test manual should be complete enough that persons tested will understand the task as the author intended. Essential

[Comment: For example, in a personality inventory, it may be intended that the subject give the first response that occurs to him. If so, this expectation should be made clear in the directions read by or to the subject. Directions for interest inventories should specify whether the person is to mark what things he would ideally like to do or whether he is also to consider the possibility that he would have the opportunity and ability to do them. Likewise, the directions should specify whether the person is to mark those things he would like to do and does occasionally, or only those things he would like to do and does regularly.]

C1.1.1. The directions should clearly point out such critical matters as instructions on guessing, time limits, and procedures for marking answer sheets. Essential

C1.1.2. The directions to the test administrator should include guidance for dealing with questions from examinees. Very Desirable

C1.2. If expansion or elaboration of instructions described in the test manual is permitted, the conditions under which this may be done should be clearly stated either in the form of general rules or in terms of giving numerous examples, or both. Essential

C2. Instructions should prepare the examinee for the examination: Sample material, practice use of answer sheets or punch cards, sample questions, etc., should be provided. Desirable

[Comment: The extent and nature of such material depends on expected levels of knowledge among examinees. For example, extensive practice material might be wasteful if developed for frequently tested school children and for a commonly encountered type of test; it may be very important for a novel test format to be administered to older job applicants.]

C3. The procedures for scoring the test should be presented in the test manual with a maximum of detail and clarity to reduce the likelihood of scoring error. Essential

C3.1. The test manual should furnish scoring instructions that maximize

the accuracy of scoring an objective test by outlining a procedure for checking the obtained scores for computational or clerical errors. Very Desirable

C3.2. Where subjective processes enter into the scoring of a test, evidence on the degree of agreement between independent scorings under operational conditions should be presented in the test manual. If such evidence is not provided, the manual should draw attention to scoring variations as a possible significant source of errors of measurement. Very Desirable

C3.2.1. The bases for scoring and the procedures for training scorers should be presented in the test manual in sufficient detail to permit other scorers to reach the level of agreement reported in studies of scorer agreement given in the manual. Very Desirable

C3.2.2. If persons having various degrees of supervised training are expected to score the test, studies of the interscorer agreement at each skill level should be presented in the test manual. Desirable

C3.3. If the test is designed to use more than one method for the examinee's recording of his reponses, such as hand-scored answer sheets, or entering of responses in the test booklet, the test manual should report data on the degree to which results from these methods are interchangeable. Essential

[Comment: The different amounts of time required for responding to items in forms adapted to different scoring methods may affect the reliability or validity of the test or the applicability of the test norms.]

C3.4. If an unusual or complicated scoring system is used, the test manual should indicate the approximate amount of time required to score the test. Desirable

C3.5. "Correction-for-guessing" formulas should be used with multiple-choice and true-false items when the test is speeded. Desirable

D. Norms and Scales

Interpretations of test scores traditionally have been *norm referenced*; that is, an individuals score is interpreted in terms of comparisons with scores made by other individuals. Alternative interpretations are possible.- *Content-referenced* interpretations are those where the score is directly interpreted in terms of performance at each point on the achievement continuum being measured. *Criterion-referenced* interpretations are those where the score is directly interpreted in terms of performance at any given point on the continuum of an *external* variable. An external criterion variable might be grade averages or levels of job performance.[2]

[2]Current usage in educational measurement commonly refers to "criterion-referenced" interpretations for *both* alternatives to interpretations requiring norms. The different meanings of the word "criterion," however, produce some confusion; some measurement specialists have therefore turned to the term "content referenced" and this usage is adopted here. The word "criterion," as it is used in the phrase "criterion-related" validity (that is, an external variable) has suggested a similar but distinguishable alternative to normative interpretation; therefore, "content-referenced and criterion-referenced" are not interchangeable terms as used in this document.]

The standards in this section refer principally to tests intended for norm-referenced test interpretations rather than for content-referenced interpretations.

D1. Norms should be published in the test manual at the time of release of the test for operational use. Essential

D1.1. Norms should be established even for a test developed only for local use or only for predictive purposes. Desirable

[Comment: It is sometimes forgotten that norms tables provide information useful for purposes other than comparing one individual with group data. For example, a test user can derive information from a normative table about the score levels at which the discrimination power of the measurement is good or poor.]

D1.2. Even though a test is expected to be used primarily with local norms, the test manual should nevertheless provide normative data to aid the interpreter who lacks local norms. Very Desirable

[Comment: The manual for one instrument designed to measure employee aptitude stresses the value of local norms but also includes norms based on a wide variety of occupational and educational classifications.]

D2. Norms presented in the test manual should refer to defined and clearly described populations. These populations should be the groups with whom users of the test will ordinarily wish to compare the persons tested. Essential

[Comment: It should be noted that "populations" are plural; in nearly all instances of tests developed for other than purely local use, the user needs to know the applicability of the test to different groups. For tests developed with a view to widespread use in schools or industry, information is needed about differences or similarities of normative data for appropriate subgroups such as sex, ethnic, grade, or age groups. Users need to be alert to situations when norms are less extensive for one group than another.

For example, the manual for an occupational interest inventory, or for an aptitude test particularly useful in certain occupations, should point out that a person who has a high degree of interest or aptitude in a curriculum or occupation when compared to people in general will usually have a lower degree of interest or aptitude compared to persons actually engaged in that field. Thus, a high percentile score on a scale reflecting musical interest, in which the examinee is compared with people in general, may be equivalent to a low percentile where the examinee is compared with professional musicians.]

D2.1. Care should be taken to avoid misleading impressions about the generality of normative data. Essential

[Comment: Truly representative national norms, for example, are rarely if ever obtained; normative data collected from people or schools with specific characteristics, however, are frequently used as if they were taken from a representative national group. Thus, we have test users who may say that an examinee's performance is at a "tenth-grade reading level," without qualification when the norms are in fact

ations" are plural; in nearly all es tests developed for other urely local use, the user needs to he applicability of the test to dif- groups. For tests developed with to widespread use in schools or ry, information is needed about nces or similarities of normative r appropriate subgroups such as nnic, grade, or age groups. Users be alert to situations when norms s extensive for one group than r.

example, the manual for an oc- onal interest inventory, or for an le test particularly useful in cer- cupations, should point out that on who has a high degree of in- or aptitude in a curriculum or oc- on when compared to people in l will usually have a lower degree rest or aptitude compared to per- actually engaged in that field. a high percentile score on a scale ing musical interest, in which the nee is compared with people in l, may be equivalent to a low per- which the examinee is compared rofessional musicians.]

1. Care should be taken to avoid ding impressions about the lity of normative data. Essential

nment: Truly representative na- norms, for example, are rarely if obtained; normative data ed from people or schools with c characteristics, however, are ntly used as if they were taken a representative national group. we have test users who may say n examinee's performance is at a -grade reading level," without cation when the norms are in fact obtained only from superior schools voluntarily participating in the test research. It is an error of interpretation to assume that the norms of the volunteer group of schools apply to schools in general; the incidence of such erros may be reduced by manuals that clearly define the characteristics of the normative populations.]

D2.1.1. The test manual should report the method of sampling from the population of examinees and should discuss any probable bias in this sampling procedure. Essential

D2.1.2. Norms reported in any test manual should be based on well-planned samplings rather than on data collected primarily because it is readily available. Any deviations from the plan should be reported along with descriptions of actions taken or not taken with respect to them. Essential

[Comment: Occupational and educational test norms have often been based on scattered groups of test papers, for authors sometimes have requested that all users mail in results for use in subsequent reports of norms. Distributions so obtained are subject to unknown degrees and types of biases. Hence, the methods of obtaining such samples should be clearly described.]

D2.1.3. In addition to reporting the number of individuals in a set of normative data, the manual should also report the number of sampling units from which those individuals were drawn along with the numbers of individuals in each unit. Essential

D2.2. The description of the norms group in the test manual should be complete enough so that the user can judge its appropriateness for his use. The description should include number of cases, classified by one or more of such relevant variables as ethnic mix, socioeconomic level, age, sex, locale, and educational status. If cluster sampling is employed, the description of the norms group should state the number of separate groups tested. Essential

[Comment: Manuals often use too gross a classification system in describing their normative data. For example, the manual for one employee aptitude test provides a variety of normative data for many occupational and educational groupings. However, the lack of information as to sex, ethnic origins, age, education, and experience levels within these groupings considerably reduces the usefulness of the norms.]

D2.2.1. The populations upon which the psychometric properties of a test were determined and for which normative data are available should be *clearly and prominently described* in the manual. Any accompanying report forms should provide space for identifying the normative groups used in interpreting the scores. Essential

[Comment: The intent of this standard is to provide a warning to consumers (users and examinees) against unwarranted interpretations. If a standard report form results in percentile-rank or standard-score interpretations by consistently using the same normative population, the definition of that population, with an indication of the time period of data collection, would be sufficient.]

D2.3. If the sample on which norms are based is small or otherwise undependable, the user should be cautioned explicitly in the test manual regarding the possible magnitude of errors arising in the interpretation of scores. Very Desirable

D2.4. Norms on subtests or groups of test items should be reported in the test manual only if the validity and reliability of such subtests or groups of items are also indicated. Essential

[Comment: The test user is justified in assuming that, when norms are given for part of a test, the author implies their usefulness for interpreting performance. The reliability and validities of such scores should be reported.]

D2.5. The significant aspects of conditions under which normative data were obtained should be reported in the test manual. Essential

[Comment: Some tests are standardized on job-applicant groups, others on groups that have requested vocational guidance, and, still others, on groups that realized that they were experimental subjects. While precise description of levels is probably not always possible, motivation for taking tests, test-taking attitudes, abilities, and personality characteristics often differ within these groups and from group to group.]

D3. In reporting norms, test manuals should use percentiles for one or more appropriate reference groups or standard scores for which the basis is clearly set forth; any exceptional type of score or unit should be explained and justified. Measures of central tendency and variability always should be reported. Essential

D3.1. In the case of tests used for prediction, expectancy tables or experience tables translating obtained scores into probabilities of success or into proficiency levels should be included whenever possible. Desirable

D4. Local norms are more important for many uses of tests than are published norms. A test manual should suggest using local norms in such situations. Very Desirable

D5. Derived scales used for reporting scores should be carefully described in the test manual to increase the likelihood of accurate interpretation of scores by both the test interpreter and the examinee. Essential

[Comment: It would be helpful if the number of kinds of derived scales could be reduced to a few with which testers can become familiar. The present variety makes description necessary in each manual. In part the problem is that many different systems are now used that have no logical advantage over others; some may have outlived their usefulness. New scaling methods may be used in attempts to overcome presumed difficulties with older ones. The variety of scales for reporting test scores can create confusion and misinterpretation unless the scales recommended for a given test are clearly and fully explained.]

D5.1. Derivation of any scale from normative data should be clearly and unambiguously described in terms likely to prevent user misinterpretations or overgeneralization. Essential

[Comment: Derived scores can be very useful for drawing inferences. Too often, however, they are treated as if they had absolute meaning independent of a particular test or normative population. An example is the IQ: it is often simply a standard score, but it is frequently reified and interpreted as representing an unchanging and unchangeable characteristic of the person tested. Grade-equivalent scores or even percentile ranks may also be misinterpreted as absolute entities unless the manual makes clear the reference group on which they were based.]

D5.2. When standard scores are used, the system should be consistent with the purposes for which the test is intended and should be described in detail in the test manual. The reasons for choosing one scale in preference to another should also be made clear in the manual. Very Desirable

D5.2.1. The manual should specify whether standard scores are linear transformations of raw scores or are normalized. Essential

D5.2.2. The choice of a standard scale should be based upon either the standard error of measurement of raw scores or on some other basis that is clearly defined. Desirable

[Comment: There are many standard-score scales in use. The scale for reporting scores on one widely used test is so designed that each unit of the scale is equal to about one thirtieth of the overall standard error of measurement; a different test used for similar purposes is scaled so that one unit is equal to about one third the overall standard error of measurement. The former scale suggests a greater degree of precision than the latter, but this implication is unwarranted.]

D5.2.3. Interpretive scores that lend themselves to gross misinterpretations, such as mental-age or grade-equivalent scores, should be abandoned or their use discouraged. Very Desirable

[Comment: When, despite this recommendation, such scores are included in a manual, their relationship to standard scores or percentile ranks, within each category and within an appropriate norm group, should also be provided in tabular form. For example, the table might show, in addition to a grade-equivalent score, the corresponding percentile rank within the examinee's own age or grade level for each raw score. At the high school level, norms within courses (for example, second-year Spanish) may be more appropriate than norms within grades.]

D5.3. When it is suggested in the test manual that percentile ranks are to be plotted on a profile sheet, the profile sheet should be based on the normal probability scale or some other appropriate nonlinear transformation. Very Desirable

D5.4. Normative data should be provided in a form that emphasizes the fallibility of an obtained score. Very Desirable

[Comment: Some publishers provide norms showing ranges of standard scores or percentile ranks that have designated probability levels including the true score. A norms table might show for each raw score, not only the

associated standard score or percentile rank but also the values for raw scores at plus and at minus one standard error of measurement for each raw score.

D6. If scales are revised, new forms added, or other changes made, the revised test manual should provide tables of equivalence between the new and the old forms. This provision is particularly important in cases where data are recorded on cumulative records. Desirable

[Comment: New forms of a test should be equated to *recently determined* standard-score scales of other forms, in order that the user may be confident that the scores furnished by the new forms are comparable with those of earlier forms.]

D6.1. When a new form is equated with an older form of a test, the revised manual should describe the content of both old and new forms and the nature of the norm group for each form. Essential

[Comment: Changes in knowledge, technology, or curricula may require that new editions of a test differ in important respects from earlier editions, and the demand for continuity may require that the scales for reporting scores be equated. There is some doubt as to whether meaningful comparability of scores is possible with changed content, however, and a user should be able to evaluate claims of equivalency in terms of the kinds of content changes that have occurred.]

D6.2. The manual should describe the method used to establish equivalent or comparable scores and should include an assessment of the accuracy of the equating procedure. Very Desirable

D7. Where it is expected that a test will be used to assess groups rather than individuals (i.e., for schools or programs), normative data based on group summary statistics should be provided. Essential

[Comment: For example, it is inappropriate to evaluate schools by using norms developed for the evaluation of individuals. It is also inappropriate to compute group means for nonlinear scales such as percentile ranks derived for individual norms.

important respects from earlier editions, and the demand for continuity may require that the scales for reporting scores be equated. There is some doubt as to whether meaningful comparability of scores is possible with changed content, however, and a user should be able to evaluate claims of equivalency in terms of the kinds of content changes that have occurred.]

D6.2. The manual should describe the method used to establish equivalent or comparable scores and should include an assessment of the accuracy of the equating procedure. Very Desirable

D6.3. Where it is expected that a test will be used to assess groups rather than individuals (i.e., for schools or programs), normative data based on group summary statistics should be provided. Essential

[Comment: For example, it is inappropriate to evaluate schools by using norms developed for the evaluation of individuals. It is also inappropriate to compute group means for nonlinear scales such as percentile ranks derived for individual norms.

STANDARDS FOR REPORTS OF RESEARCH ON RELIABILITY AND VALIDITY

A test developer must provide evidence of the reliability and validity of his test; it is usually reported in the test manual. Many test users should do similar research on their own application of the test. Their reports often differ from those in test manuals by being more detailed or more specific to a particular problem, or by validating test batteries rather than individual tests. Despite such differences, the standards of research, and of research reporting, should be generally similar in the two situations.

E. Validity

Questions of validity are questions of what may properly be inferred from a test score; validity refers to the appropriateness of inferences from test scores or other forms of assessment. The many types of validity questions can, for convenience, be reduced to two: (*a*) What can be inferred about what is being measured by the test? (*b*) What can be inferred about other behavior?

The first question inquires into the intrinsic nature of the measurement itself. The measuring instrument is an operational definition of a specified domain of skill or knowledge, or of a trait, of interest to the test developer or user. The essential problem in this context is to reach some conclusion as to how faithfully the scores represent that domain, and it is appropriate to speak of the validity of the measurement.

The second question inquires into the usefulness of the measurement as an indicator of some other variable as a predictor of behavior. In this context, the essential problem is to reach some conclusion about how well scores on the test are related to some other performance, and it is appropriate to speak of the closeness of the relationship.

The two questions are not necessarily independent. For example, where the test is a sample of the "other behavior," the answer is the same for either question. Moreover, answers to both questions may require a knowledge of the interrelationships between the test scores and other variables. A thorough understanding of validity may require many investigations. The investigative processes of gathering or evaluating the necessary data are called validation. There are various methods of validation, and all, in a fundamental sense, require a definition of what is to be inferred from the scores and data to show that there is an acceptable basis for such inferences.

It is important to note that validity is itself inferred, not measured. Validity coefficients may be presented in a manual, but validity for a particular aspect of test use is inferred from this collection of coefficients. It is, therefore, something that is *judged* as adequate, or marginal, or unsatisfactory.

The kinds of validity depend upon the kinds of inferences one might wish

to draw from test scores. Four interdependent kinds of inferential interpretation are traditionally described to summarize most test use: the *criterion-related* validities (*predictive* and *concurrent*); *content* validity; and *construct* validity.[3] (So-called "face" validity, the mere appearance of validity, is not an acceptable basis for interpretive interferences from test scores.)

These aspects of validity can be discussed independently, but only for convenience. They are interrelated operationally and logically; only rarely is one of them alone important in a particular situation. A thorough study of a test may often involve information about all types of validity. In developing or choosing a test for prediction, one should first postulate the constructs likely to provide a basis for useful prediction of the variable of interest; the measures chosen should have adequate construct validity. The content universe from which items are sampled may also be an important early step in producing a predictive test, in evaluating a test considered for use as a predictor, or in developing the criterion measure to be predicted. Even if the accuracy of prediction is good, information about construct validity may make a test more useful. To evaluate construct validity, all knowledge regarding validity is relevant. A reading comprehension test, for example, may be used and validated for all three types of inference: how well it predicts future academic performance, how well it samples a defined content area of material to read, and how well it measures the construct of comprehension.

Criterion-Related Validities

Criterion-related validities apply when one wishes to infer from a test score an individual's most probable standing on some other variable called a criterion. Statements of predictive validity indicate the extent to which an individual's future level on the criterion can be predicted from a knowledge of prior test performance; statements of concurrent validity indicate the extent to which the test may be used to estimate an individual's present standing on the criterion. The distinction is important. Predictive validity involves a time interval during which something may happen (e.g., people are trained, or gain experience, or are subjected to some treatment). Concurrent validity reflects only the status quo at a particular time. Under appropriate circumstances, data obtained in a concurrent study may be used to estimate the predictive validity of a test. However, concurrent validity should not be used as a substitute for predictive validity without an appropriate supporting rationale.

For many test uses, such as for selection decisions or assignment to treatment, predictive validity provides the appropriate model for evaluating the use of a test or test battery. In employment testing, for example, use of

[3]Many other terms have been used. Examples include synthetic validity, convergent validity, job-analytic validity, rational validity, and factorial validity. In general, such terms refer to specific procedures for evaluating validity rather than to new kinds of interpretive inferences. Any specially-named procedures, including these examples, should meet the standards of investigation contained in this section. These standards apply generally to the various statistics or procedures that might be used in support of one or more classes of inferences from test scores.

evaluate construct validity, all knowledge regarding validity is relevant. A reading comprehension test, for example, may be used and validated for all three types of inference: how well it predicts future academic performance, how well it samples a defined content area of material to read, and how well it measures the construct of comprehension.

Criterion-Related Validities

Criterion-related validities apply when one wishes to infer from a test score an individual's most probable standing on some other variable called a criterion. Statements of predictive validity indicate the extent to which an individual's future level on the criterion can be predicted from a knowledge of prior test performance; statements of concurrent validity indicate the extent to which the test may be used to estimate an individual's present standing on the criterion. The distinction is important. Predictive validity involves a time interval during which something may happen (e.g., people are trained, or gain experience, or are subjected to some treatment). Concurrent validity reflects only the status quo at a particular time. Under appropriate circumstances, data obtained in a concurrent study may be used to estimate the predictive validity of a test. However, concurrent validity should not be used as a substitute for predictive validity without an appropriate supporting rationale.

For many test uses, such as for selection decisions or assignment to treatments, predictive validity provides the appropriate model for evaluating the use of a test or test battery. In employment testing, for example, use of

any procedure implies prediction to some degree. Whether one uses a carefully developed test or casual judgments of interviewers, their use for selection purposes assumes that applicants who obtain high scores will become better employees than applicants who obtain low scores.

Other forms of validity are not substitutes for criterion-related validity. In choosing a test to select people for a job, for example, an abundance of evidence of the construct validity of a test of flexibility in divergent thinking, or of the content validity of a test of elementary calculus, is of no predictive value without reason to believe that flexibility of thinking or knowledge of calculus aids performance on that job. The *model* of predictive validity should guide thinking about validity in such applications even where circumstances preclude an actual criterion-related validation study. Whatever other validity information a manual may include, one or more studies of criterion-related validity must be included for any test developed for prediction and for many tests intended for diagnosis; otherwise, such tests can only be regarded as experimental.

Many factors may make a single, obtained validity coefficient questionable. First, the conditions of a validation study are never exactly repeated. Rapidly changing conditions may limit the usefulness of a predictive study. The logic of predictive validation assumes that conditions existing at the start of the time sequence will exist again after the study is completed.

Second, the logic of criterion-related validity assumes that the criterion possesses validity. All too often, tests are validated against any available criterion with no corresponding investigation of the criterion itself. The merit of a criterion-related validity study depends on the appropriateness and quality of the criterion measure chosen. In applied research, the criterion should be chosen with reference to the problem at hand, and the test or other assessment technique should be chosen with reference to the criterion. If the study is done primarily to enhance understanding of what a test measures, criteria should be selected in terms of beliefs about the nature of the construct reflected by the test scores. In either case, the adequacy of the study depends on the adequacy of the criterion. Criterion-related validity studies based on the "criterion at hand," chosen more for availability than for a place in a carefully reasoned hypothesis, are to be deplored.

Third, the logic of criterion-related validity assumes that the sample is truly representative of the population for which the later inferences are to be drawn. In practice, samples are often nonrepresentative because of, for example, restricted range, preselection, or attrition before a predictive study can be completed.

Fourth, in many practical situations validity studies cannot be done with adequate numbers of cases, and the investigators must do the best they can with the data at hand. It may be better to try to investigate criterion-related validity, even if imperfectly, than to accept totally untested hypotheses. However, "doing something" is not necessarily better than doing nothing; the results of an inadequate study may be quite misleading. Results of

validation studies with severely restricted ranges or small Ns are especially open to question.

Content Validity

Evidence of content validity is required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent. Content validity is most commonly evaluated for tests of skill or knowledge; it may also be appropriate to inquire into the content validities of personality inventories, behavior checklists, or measures of various aptitudes. The present discussion will be directed toward the more typical case of achievement testing.

To demonstrate the content validity of a set of test scores, one must show that the behaviors demonstrated in testing constitute a representative sample of behaviors to be exhibited in a desired performance domain. Definitions of the performance domain, the users' objectives, and the method of sampling are critical to claims of content validity. An investigation of content validity requires that the test developer or test user specify his objectives and carefully define the performance domain in light of those objectives. The definition should ordinarily specify the results of learning rather than the processes by which learning is either sufficiently detailed and organized to show the degree to which component tasks make up the total domain.

Definition of the performance domain is relatively simple where it is finite and unambiguous, as in a simple test of addition for elementary-school use. Depending upon instructional objectives, the performance domain might be defined as *all* addition problems of three to five single-integer addends. The total number of problems and the relative frequency of occurrence of specific integers or pairs of integers within that total are known, and the representativeness of any sample of such problems can be easily judged.

If a test is used to estimate achievement in American history in Grade 12, the performance domain is less objectively defined. Given agreement on instructional objectives, it could be defined in terms of the types and quantities of the skills, facts, and concepts of American history, as determined by the pooled judgments of authorities, experienced teachers, and competent curriculum makers in that field. A definition of the total universe might well be tempered by the specific instructional objectives accepted by the panel. A definition appropriate for evaluation of performance at the end of the year of study would differ from the definition appropriate for developing an examination over knowledge of the colonial period. Within such limits, the performance domain requires definition so carefully detailed that rules for item writing will assure appropriate representation of all facets of the definition. It should be noted that an achievement test so constructed would not necessarily constitute a representative sample of the skills, facts, and concepts taught by any particular teacher during any particular year. Consequently, a definition of the performance domain of interest must always be provided by a test user so that the content of a test may be checked against an appropriate task universe.

objectives, the performance domain might be defined as *all* addition problems of three to five single-integer addends. The total number of problems and the relative frequency of occurrence of specific integers or pairs of integers within that total are known, and the representativeness of any sample of such problems can be easily judged.

When a test is used to estimate achievement in American history in Grade 12, the performance domain is less objectively defined. Given agreement on instructional objectives, it could be defined in terms of the types and quantities of the skills, facts, and concepts of American history, as determined by the pooled judgments of authorities, experienced teachers, and competent curriculum makers in that field. A definition of the total universe might well be tempered by the specific instructional objectives accepted by the panel. A definition appropriate for evaluation of performance at the end of the year of study would differ from the definition appropriate for developing an examination over knowledge of the colonial period. Within such limits, the performance domain requires definition so carefully detailed that rules for item writing will assure appropriate representation of all facets of the definition. It should be noted that an achievement test so constructed would not necessarily constitute a representative sample of the skills, facts, and concepts taught by any particular teacher during any particular year. Consequently, a definition of the performance domain of interest must always be provided by a test user so that the content of a test may be checked against an appropriate task universe.

It is appropriate to inquire into the content validity of many employment tests. Examples would include tests of typing skill, driving ability, or knowledge of certain regulatory laws. The performance domain for published tests might be defined by the pooled judgments of job designers, incumbents, and supervisors. Test users might define the performance domain of interest to them in terms of judgments of similar people in their own organizations or, preferably, in terms of appropriately detailed and comprehensive job analyses. The question of objectives would again enter into the definition; unless only fully trained and experienced people are to be hired, applicants cannot be expected to demonstrate proficiency in all facets of a job. The performance domain would need definition in terms of the objectives of measurement, restricted perhaps only to critical, most frequent, or prerequisite work behaviors.

It should be clear that content validity is quite different from face validity. Content validity is determined by a set of operations, and one evaluates content validity by the thoroughness and care with which these operations have been conducted. In contrast, face validity is a judgment that the requirements of a test merely *appear* to be relevant. The writing of items in terms used in a particular job or by a particular subgroup of the population may give an appearance of relevance while contributing nothing to content validity or indeed to any other useful validity information (although such items may serve a useful public-relations function).

In defining the content universe, a test developer or user is accountable for the adequacy of his definition. An employer cannot justify an employment test on grounds of content validity if he cannot demonstrate that the content universe includes all, or nearly all, important parts of the job.

Construct Validity

A psychological construct is an idea developed or "constructed" as a work of informed, scientific imagination; that is, it is a theoretical idea developed to explain and to organize some aspects of existing knowledge. Terms such as "anxiety," "clerical aptitude," or "reading readiness" refer to such constructs, but the construct is much more than the label; it is a dimension understood or inferred from its network of interrelationships.[4] It may be necessary to postulate several different constructs to account for the variance in any given set of test scores. Moreover, different constructs may be required to account for the variance in different tests of the same general type, or a given test may provide evidence relating to several constructs. For example, given proper evidence, scores on vocabulary tests might be used to infer (*a*) the level of present vocabulary; (*b*) the existence of pathology, interests, or values; or (*c*) intellectual capacity.

Construct validity is implied when one evaluates a test or other set of operations in light of the specified construct. Judgments of construct

[4]This is an admittedly restricted statement of the nature of scientific constructs, which may include entities as well as dimensions. Constructs of interest in the present context are, however, primarily quantitative.

validity are useful in efforts to improve measures for the scientific study of a construct. They are also useful when a test developer or test user wishes to learn more about the psychological qualities being measured by a test than can be learned from a single criterion-related validity coeffficient.

Evidence of construct validity is not found in a single study; rather, judgments of construct validity are based upon an accumulation of research results. In obtaining the information needed to establish construct validity, the investigator begins by formulating hypotheses about the characteristics of those who have high scores on the test in contrast to those who have low scores. Taken together, such hypotheses form at least a tentative theory about the nature of the construct the test is believed to be measuring. In a full investigation, the test may be the dependent variable in some studies and the independent variable in others. Some hypotheses may be "counterhypotheses" suggested by competing interpretations or theories.

Such hypotheses or theoretical formulations lead to certain predictions about how people at different score levels on the test will behave on certain other tests or in certain defined situations. If the investigator's theory about what the test measures is essentially correct, most of his predictions should be confirmed. If they are not, he may revise his definition of the construct, or he may revise the test to make it a better measure of the construct he had in mind. Through the process of successive verification, modification, or elimination of hypotheses, the investigator increases his understanding of the qualities measured by the test. Through the process of confirmation or disconfirmation, test revision, and new research on the revised instrument, he improves the usefulness of the test as a measure of a construct.

It is important to note in this that the investigation of construct validity refers to a specific test and not necessarily to any other test given the same label.

Evidence of construct validity may also be inferred from the procedures followed in developing a test. For example, in a measure of mechanical interest, a double item analysis may be used to reduce the effect of verbal ability. A preliminary item analysis might be done using a standard verbal-comprehension test as an external criterion. Those items with a very low discrimination index in this analysis could then be subjected to a second item analysis, a conventional internal-consistency analysis. Only those items with a low discrimination index in the first analysis and a high discrimination index in the second analysis would be included in the final item pool.

Although evidence of construct validity may be developed on the basis of a series of criterion-related studies, it is important to note that evidence of the construct validity of a test is *not* adequate evidence of the usefulness of the construct in specific further hypotheses. In the selection of salespersons, for example, it is often hypothesized that success is a function of sociability. If one has a measure of sociability with generally acceptable evidence of its validity as a measure of that construct, he may expect to find it

heses, the investigator increases his understanding of the qualities measured by the test. Through the process of confirmation or disconfirmation, test revision, and new research on the revised instrument, he improves the usefulness of the test as a measure of a construct.

It is important to note in this that the investigation of construct validity refers to a specific test and not necessarily to any other test given the same label.

Evidence of construct validity may also be inferred from the procedures followed in developing a test. For example, in a measure of mechanical interest, a double item analysis may be used to reduce the effect of verbal ability. A preliminary item analysis might be done using a standard verbal-comprehension test as an external criterion. Those items with a very low discrimination index in this analysis would then be subjected to a second item analysis, a conventional internal-consistency analysis. Only those items with a low discrimination index in the first analysis and a high discrimination index in the second analysis would be included in the final item pool.

Although evidence of construct validity may be developed on the basis of a series of criterion-related studies, it is important to note that evidence of the construct validity of a test is *not* adequate evidence of the usefulness of the construct in specific further hypotheses. In the selection of salespersons, for example, it is often hypothesized that success is a function of sociability. If one has a measure of sociability with generally acceptable evidence of its validity as a measure of that construct, he may expect to find it

useful as a predictor of sales success; perhaps some of the evidence of the construct validity of that measure came, in fact, from confirmation of such an expectation. However, the test may have no predictive validity against the criterion of success in an engineering sales job. In such a case it is not the construct validity of the sociability measure that is to be questioned.

General Principles

A test developer, or anyone who conducts validation research, should provide as much validity information as possible so the user can evaluate the test or the research for his own purposes. A test manual can provide evidence that will enable the user to evaluate the appropriateness of the item content, to determine whether the test is an acceptable measure of a specified construct, and to decide whether the test has provided useful predictive validities in situations similar to his own. An adequate research report can help the user decide whether to go ahead with the use of the test or to seek another predictor.

E1. A manual or research report should present the evidence of validity for each type of inference for which use of the test is recommended. If validity for some suggested interpretation has not been investigated, that fact should be made clear. Essential

[Comment: Validation studies are a part of the process of test development; test users expect them to be reported in detail by the developer, preferably in the manual itself. At the very least, the manual should summarize competent research reported elsewhere, either by the test developer or by others. Preferably, the manual will report on individual studies and provide summaries of validity data for various kinds of interpretations or inferences.]

E1.1. Statements about validity should refer to the validity of particular interpretations or of particular types of decisions. Essential

[Comment: It is incorrect to use the unqualified phrase "the validity of the test." *No test is valid for all purposes or in all situations or for all groups of individuals.* Any study of test validity is pertinent to only a few of the possible uses of or inferences from the test scores.

If the test is likely to be used incorrectly for certain areas of decision, the manual should include specific warnings. For example, the manual for a writing-skills test stated that the test apparently was not sufficiently difficult to discriminate among students "at colleges that have selective admissions."]

E1.2. Wherever interpretation of subscores, score differences, or profiles is suggested, the evidence justifying such interpretation should be made explicit. (See also B5.5.) Essential

E1.2.1. If the manual suggests that the user consider an individual's responses to specific items as a basis for assessment, it should either present evidence supporting this use or call attention to the absence of such data. The manual should warn the reader that inferences based on responses to single items are subject to extreme error. Hence, they should be used only to

direct further inquiry, perhaps in a counseling interview. Essential

E1.3. To insure the continued correct interpretation of scores, the validity of suggested interpretations should be rechecked periodically; test developers should report results in subsequent editions of the manual. Very Desirable

[Comment: Job duties, conditions of work, and the types of individuals entering an occupation often change materially with the passage of time. Similarly, the meanings of clinical categories, the nature of therapeutic treatment, and the objectives of academic programs change. The difficulty and psychological meaning of test items will also change. Hence, the reader should be in a position to judge the extent to which tests are obsolete.]

E1.3.1. If factors that may affect test performance or the validity of a suggested test interpretation have changed, and validity studies have not been repeated for the changed conditions, the test should be withdrawn from general sale and distributed, if at all, only to persons who will conduct their own validity studies. Very Desirable

[Comment: It should be noted that no specific time interval is mentioned. Test developers and publishers should know the relevant conditions and should be able and willing to obtain new validity information when such conditions have changed. It is not necessary to repeat every part of the validation; what is needed is a repetition of those studies most likely to have been rendered obsolete. In the case of some inventories and biographical-data forms, scoring keys should be reevaluated after relatively brief periods of time.]

E1.4. Correlations of item scores with total scores on the test in which the item is included (or a parallel form of that test) may be presented as item-discrimination coefficients, but they should not be presented or used as item-validity coefficients. Essential

[Comment: Item-discrimination coefficients are useful in reasoning about construct validity, and such information is appropriately included in a manual. However, they are indicators of internal consistency, not of validity.]

E2. A test user is responsible for marshalling the evidence in support of his claims of validity and reliability. The use of test scores in decision rules should be supported by evidence. Essential

[Comment: It is a basic responsibility of a test user to read, understand, and evaluate the manual, the research, and the literature to show the appropriateness of the test for the intended use. A large-scale user may have the added responsibility for empirical research bearing on his claims of test validity. Evidence of validity is needed for *all* bases for decision, not merely those that are easy to study. It is a peculiar paradox that many employers and schools are abandoning the use of standardized tests and are turning instead to casual assessment techniques likely to be less valid. Many employers use procedures with no validity, or biased selection

ntor and biographical-data s, ring keys should be reevalu- after relatively brief periods of -].

.4. Correlations of item scores total scores on the test in which the is included (or a parallel form of tes may be presented as item- imination coefficients, but they ld not be presented or used as -val ty coefficients. Essential

om nt: Item-discrimination co- ents are useful in reasoning about truct validity, and such rma n is appropriately included mal al. However, they are indica- of internal consistency, not of ity.

2. test user is responsible for shalling the evidence in support of cl s of validity and reliability. us f test scores in decision rules ld be supported by evidence. ential

Com ent: It is a basic responsibility tes ser to read, understand, and uate the manual, the research, and rature to show the rop teness of the test for the nded use. A large-scale user may e the added responsibility for pirio research bearing on his ns f test validity. Evidence of dity is needed for *all* bases for sion not merely those that are easy tud it is a peculiar paradox that y mployers and schools are ndoning the use of standardized s an are turning instead to casual essn t techniques likely to be less d. Many employers use procedures n no validity, or biased selection procedures of unknown validity rather than objective procedures for which evidence of validity could have been assembled.]

E2.1. Test users are responsible for gathering data on the validity and reliability of their assessment techniques. Very Desirable

[Comment: For many individual test users, this may be a nearly impossible requirement. It would seem, however, that a test user has an obligation to gather data, at least on an informal basis, in an effort to evaluate his work. In even the most difficult circumstances, a test user can be alert to data suggesting possible lack of validity.]

E2.2. If a user wants to use a test in a situation for which the use of the test has not been previously validated, or for which there is no supported claim for validity, he is responsible for validation. Very Desirable

[Comment: He who makes the claim for validity is responsible for providing the evidence. Evidence of validity sufficient for test use may often be obtained in a well-documented manual. If the test user wishes to claim that the validity generalizes beyond the evidence for the kinds of situations reported in the manual, it is his responsibility to demonstrate it.]

E2.3. When a test user plans to make a substantial change in test format, instructions, language, or content, he should revalidate the use of the tests for the changed conditions. Essential

Criterion-Related Validity

E3. All measures of criteria should be described completely and accurately. The manual or research report should comment on the adequacy of a criterion. Whenever feasible, it should draw attention to significant aspects of performance that the criterion measure does not reflect and to irrelevant factors likely to affect it. Essential

[Comment: Desirable practices are illustrated in a manual for a test designed to measure abstract intelligence. Several validity studies relating this instrument to criteria are reported, some involving concurrent measures and others involving predictions over periods of time. Limitations of the studies are recognized, and it is stated that "no one criterion is uniquely appropriate." The value of local norms is stressed, and an example of a local expectancy table is provided.

In the case of interest measures, it is sometimes not made clear whether the criterion indicates satisfaction, success, or merely continuance in the activity under examination. When criterion groups include people in a given occupation and when a comparison of such groups is made to people in general, the manual should point out the distinction between working in an occupation and success in it or satisfaction with it.]

E3.1. When the validity of a test for predicting occupational performance is reported, the manual should describe the duties of the workers as well as give their job titles. Very Desirable

[Comment: The principle is that information should be given from which the reader can make judgments of the relevance of the criterion. The

description of a criterion is often incomplete without such information.]

E3.1.1. Where a wide range of duties is subsumed under a given occupational label, the test user should be warned against assuming that only one pattern of interests or abilities is compatible with the occupation. Essential

E4. A criterion measure should itself be studied for evidence of validity and that evidence should be presented in the manual or report. Very Desirable

[Comment: Criterion measures are forms of assessment and are subject to the same standards governing the development and use of any assessment technique. For many employment and educational purposes, the ideal criterion may be an achievement test or work sample judged acceptable in terms of content validity. Supervisory or instructor ratings are more common but may be questioned in terms of construct validity. For example, a rating of proficiency may be defined to include elements of both speed and accuracy, but to exclude elements of dependability. A judgment of acceptable construct validity might be based on evidence of high correlations of the ratings with production data or work samples and of independence from seniority or attendance data.]

E4.1. Particular attention should be given to potential sources of criterion contamination; results of investigations of contamination should be reported. Essential

[Comment: Results of such investigations are often ambiguous, and readers should be warned of this fact. For example, an investigation of possible sex differences in criterion ratings might show significant differences between men and women. That fact in itself, however, is not sufficient evidence of criterion contamination; it might reflect actual sex differences in performance.]

E4.1.1. The criterion score should be determined independently of test scores. The manual should describe precautions taken to avoid contamination of the criterion or should warn the reader of possible contamination. Essential

[Comment: When the criterion is based on judgment, the manual should state whether the test data were available to the judge or were capable of influencing the judgments in any other way. If the test data could have influenced the criterion rating, the user should be warned that the reported validities are likely to be spuriously high.]

E4.2. The basis for judgments of criterion relevance should be clearly set forth. Essential

E.4.3. Criterion-related validation should ordinarily consider more than a single global criterion. Very Desirable

[Comment: In most situations where decision rules based on testing are worthwhile, performance falls along many dimensions that may be independent. Combining unrelated aspects of behavior into a single composite criterion may obscure important relationships and reduce a test user's opportunity to identify and understand valid test interpretations.

A problem exists in that single

For example, an investigation of possible sex differences in criterion ratings might show significant differences between men and women. That fact in itself, however, is not sufficient evidence of criterion contamination, it might reflect actual sex differences in performance.]

E4.1.1. The criterion score should be determined independently of test scores. The manual should describe precautions taken to avoid contamination of the criterion or should warn the reader of possible contamination. Essential

[Comment: When the criterion is based on judgment, the manual should state whether the test data were available to the judge or were capable of influencing the judgments in any other way. If the test data could have influenced the criterion rating, the user should be warned that the reported validities are likely to be spuriously high.]

E4.2. The basis for judgments of criterion relevance should be clearly set forth. Essential

E.4.3. Criterion-related validation should ordinarily consider more than a single global criterion. Very Desirable

[Comment: In most situations where decision rules based on testing are worthwhile, performance falls along many dimensions that may be independent. Combining unrelated aspects of behavior into a single composite criterion may obscure important relationships and reduce a test user's opportunity to identify and understand valid test interpretations.

A problem exists in that single decisions must frequently be made on multivariate bases. Nevertheless, it is preferable to find a decision rule for combining predictions than to use a decision rule for combining predictors of ambiguous validity.]

E4.4. In criterion-related validation, it is important that the criterion measure have substantial reliability. Very Desirable

[Comment: Since corrections for unreliability of a criterion can be made with generally reasonable statistical assumptions, the degree of reliability of a criterion is perhaps less important than the degree of reliability of the predictor. Nevertheless, reliability is not a trivial consideration; reasonable effort should be made to assure a level of reliability such that statistical corrections are unlikely to change interpretations markedly.]

E4.4.1. Since the criterion measure is a sample of all possible measures of the same criterion construct, reliability should be reported in terms of the agreement of that sample with other similar samples where feasible. If such evidence cannot be given, the author should make this clear and should discuss the probable extent of agreement of the sample with other samples as judged from indirect evidence. Very Desirable

[Comment: When validity is measured by agreement of the test with psychiatric judgment, for example, the degree of agreement among judges should be described. Where a published achievement test is used as a criterion measure, the form-to-form agreement or the reliability reported by the test's author may be used as a basis for evaluating the criterion, due regard being given to the effect of differences between the present sample of persons and the original sample.]

E4.4.2. When validity is appraised by agreement of test results with psychiatric diagnoses, the diagnostic terms or categories should be defined specifically and described clearly. Very Desirable

[Comment: For example, "paranoid schizophrenia, chronic" is preferable as a category to "schizophrenia." Since the types of patients included in specific diagnostic classifications depend to some extent on the point of view of the classifying psychiatrist, an amplified description of each diagnostic category used in the validity study should be presented.]

E4.4.3. When validity is appraised by the agreement of test results with psychiatric judgments, the training, experience, and professional status of the judge(s) should be stated, and the nature and extent of his contacts with the patients and other factors influencing the interaction should be reported. Very Desirable

E.5. The manual or research report should provide information on the appropriateness of or limits to the generalizability of validity information. Very Desirable

E5.1. A test manual should report evidence of validity for each type of criterion about which a recommendation is made. If validity for some

recommended interpretation has not been tested, that fact should be made clear. Essential

[Comment: This principle should not be interpreted as license to present as validity information various correlations with irrelevant variables. Data should be presented with reference to recommended interpretations, either supporting those interpretations or suggesting limitations to them.]

E5.2. For any type of prediction, a test manual should report criterion-related validities for a variety of institutions or situations. Where validity studies have been confined to a limited range of situations, the manual should remind the reader of the risks involved in generalizing to other types of situations. Essential

E5.2.1. Validity coefficients are specific to the situations in which they are obtained. If the manual is to suggest generalization of validity for prediction of a given kind of criterion construct, it must present data suggesting the limits of generalizability regarding population or sample characteristics, situational context variables, or variations in criterion measurement. Very Desirable

E5.2.2. Local collection of evidence on criterion-related validity is frequently more useful than published data. In such cases the manual should suggest appropriate emphasis on local validity studies; and test users should, where feasible, conduct such studies. Desirable

[Comment: In cases where criteria differ from one locality to another or from one institution to another, no published validity data can serve all localities. For example, the validity of a certain test for predicting grades at a college with a unique kind of curriculum may be quite different from the published validities of the same test based on a more conventional curriculum.

Some publishers have made available advice on local validation studies, including information on the preparation of expectancy tables.]

E6. The sample employed in a validity study and the conditions under which testing is done should be consistent with recommended test use and should be described sufficiently for the reader to judge its pertinence to his situation. Essential

E6.1. Any selective factor determining the composition of the validation sample should be indicated in a manual or research report. The sample should be described in terms of those variables known as thought to affect validity, such as age, sex, socioeconomic status, ethnic origin, residential region, level of education, or other demographic or psychological characteristics. Essential

[Comment: If a validity study uses patients as subjects, the diagnoses of the patients would usually be important to report. The severity of the diagnosed condition should be stated when feasible. For tests used in industry the employment status, occupational experience, and the sex and ethnic composition of the sample should be described. For tests used in educational

settings
include
any select

E6.1.1.
be obtai
same a
vocation
whom th
deviation
be descri
report. E

[Comm
tests int
generally
tested p
they a
vocation

One
standard
employed
The
differentia
did not,
the invent
school or
dence was
istering th
mining th
ment, and
between
later occ

If an
criterion
rollment
sample
include
appropri
For exa
suitable
the valid
manual
them later

differ from one locality to another or from one institution to another, no published validity data can serve all localities. For example, the validity of a certain test for predicting grades at a college with a unique kind of curriculum may be quite different from the published validities of the same test based on a more conventional curriculum.

Some publishers have made available advice on local validation studies, including information on the preparation of expectancy tables.]

E6. The sample employed in a validity study and the conditions under which testing is done should be consistent with recommended test use and should be described sufficiently for the reader to judge its pertinence to his situation. Essential

E6.1. Any selective factor determining the composition of the validation sample should be indicated in a manual or research report. The sample should be described in terms of those variables known as thought to affect validity, such as age, sex, socioeconomic status, ethnic origin, residential region, level of education, or other demographic or psychological characteristics. Essential

[Comment: If a validity study uses patients as subjects, the diagnoses of the patients would usually be important to report. The severity of the diagnosed condition should be stated when feasible. For tests used in industry the employment status, occupational experience, and the sex and ethnic composition of the sample should be described. For tests used in educational settings, relevant information may include community characteristics or any selection policies.]

E6.1.1. Evidence of validity should be obtained for subjects who are of the same age or in the same educational or vocational situation as the persons for whom the test is recommended. Any deviation from this requirement should be described in the manual or research report. Essential

[Comment: Validity information for tests intended for guidance should generally be determined on subjects tested prior to or near the time when they are making educational or vocational choices.

One interest inventory was first standardized on men currently employed in the occupation in question. The ability of these scales to differentiate between occupational groups did not, in and of itself, warrant using the inventory in the counseling of high school or college students. Better evidence was obtained later by administering the inventory to students, determining the nature of their later employment, and then establishing the relation between preoccupational score and later occupation.

If an interest inventory uses a criterion of enrollment or nonenrollment in a certain occupation, the sample used in its validation should include only the range of mental ability appropriate to the occupational group. For example, college students are not suitable subjects with whom to estimate the validity of an inventory of interest in manual skills, even though some of them later enter manual occupations.]

E6.1.2. If an ability test is to be used for educational or occupational selection, its validity should be established using subjects who are actual candidates and who are therefore ordinarily motivated to perform well. If the subjects used in a validity study are volunteers or were told that their test scores would not be used in making decisions about them, this fact should be made clear. Very Desirable

[Comment: Widespread use is made in industrial selection of the "present-employee method" of validation. Typically, this involves administering the tests to present employees who are told that their performance on the test will not influence their employment situation. The motivational difference may distort sample characteristics in that it introduces a completely new variable to the testing situation.]

E6.2. Basic statistics should be reported in describing the sample, including numbers of cases (and the reasons for any eliminated cases) and measures of central tendency and variability. A description of the distribution, perhaps with measures of skewness and kurtosis, should also be included. Very Desirable

[Comment: The smaller the number of cases in a validation sample, the less reliable the statistics. Consider, for example, the effect of the number of cases on the 95% confidence interval when the obtained correlation coefficient is .30. With 250 pairs of observations, the interval runs from .18 to .41. Where $N = 50$, the interval runs from .02 to .53; where $N = 25$, the interval runs from -.10 to +.62.

When N is very small, an obtained correlation coefficient of zero may result in the erroneous rejection of a valid test.]

E6.2.1. If the distribution of test scores in a validation sample is markedly different from the distribution of scores in the group with whom the test is ordinarily used, data based on these scores, including estimates of population parameters, should be interpreted with great caution. Reports of parameter estimates should cite the original statistics, the distribution characteristics used in making the new estimate, and the statistical procedures employed. Essential

[Comment: The reader of the manual or research report needs to evaluate any adjustments made because of atypical sample characteristics. The assumptions underlying such adjustments are frequently ignored; the resulting errors are of undetermined size and direction. Despite such difficulties, an estimate of the appropriate statistic is often needed. The reported validity coefficient, for example, should reflect the predictive power of a test in the group to which it will be applied.]

E6.2.2. Statistical corrections, such as those for restriction of range, should not be made in situations where mean performance in the sample in the validation study is so different from the mean performance of the population in which it is to be used as to suggest differences in parent populations. Very Desirable

[Comment: A validation study was conducted on a sample of applicants hired in a period when the rejection ratio was very high, resulting in a serious restriction of range. However, the passage of time brought with it changes in performance standards and in recruiting activities; the variance was greater and the mean was lower in a new sample than in the sample used in the validation study. The situational and psychometric facts together identify a change in groups more important than differences in variance alone, and a "correction" based on sample variances is therefore inappropriate.

In this situation it would be more appropriate to replicate the study with the new applicant population than to "correct" data obtained from the old population for restriction of range.]

E7. The collection of data for a validity study should follow procedures consistent with the purposes of the study. Essential

E7.1. Where feasible, a test should not be used as a basis for decision while its proposed use is being validated. Desirable

[Comment: In many practical situations, decisions must be made whether or not there is a validated basis for them. In some circumstances, high costs may necessitate the use of a test even though it has not been validated, but great caution should then be exercised. In a selection situation, for example, decisions should not be so highly selective that restriction of range makes discovery of validity impossible. Wherever possible, at least some validation research should be done before a test is put to use as a decision

red in a period when the rejection io as very high, resulting in a riou restriction of range. However, e passage of time brought with it anges in performance standards and recr ting activities; the variance was eate and the mean was lower in a w sample than in the sample used in e va dation study. The situational d sychometric facts together ntify a change in groups more portant than differences in variance ne, nd a "correction" based on mpl variances is therefore ppropriate.

In this situation it would be more pro ate to replicate the study with e new applicant population than to orrect" data obtained from the old pul on for restriction of range.]

E7. The collection of data for a idity study should follow procedures nsis t with the purposes of the dy. sential

E7.1. Where feasible, a test should t be ed as a basis for decision while pr osed use is being validated. sirable

[Comment: In many practical uati s, decisions must be made eth r not there is a validated basis them. In some circumstances, high ts y necessitate the use of a test n t gh it has not been validated, great caution should then be rcised. In a selection situation, for mp decisions should not be so hly ective that restriction of range kes discovery of validity impossible. erex r possible, at least some dat n research should be done ore test is put to use as a decision instrument, and further data should be gathered subsequently.]

E7.2. If the validity sample is made up of records accumulated haphazardly or voluntarily submitted by test users, this fact should be stated in the manual or research report, and the test user should be warned that the group is not a systematic or random sample of any specifiable population. Probable selective factors and their presumed influence on the test variable should be stated. Essential

[Comment: While it is entirely appropriate to include in the manual such phrases as "the author and publisher of this test would welcome additional data derived from its use," it is difficult to judge the quality and representativeness of most of the resulting reports.]

E7.3. In collecting data for a validity study, the person who interprets the test results should have only that information about the examinees that is ordinarily expected to be available in practical use of the test, or he should be sufficiently trained and disciplined disregard information ordinarily not available to him. If there is any possible contamination associated with prior favorable or unfavorable knowledge about the examinees, the manual should discuss its effect on the outcome of the study. Very Desirable

E7.4. The time elapsing between the test administration and the collection of criterion data should be reported in the manual. If the criterion data are collected over a period of time, beginning and ending dates should be included. Essential

E7.4.1. Validation reports should be clearly dated, with the time interval given during which the data were collected. Essential

[Comment: Validity may deteriorate over time; in employment testing, for example, changes in jobs, work aids, and in the ability levels of applicant populations tend to change the circumstances in which validity information is developed.]

E7.4.2. In general, a test user should be cautious in making long-term predictions. Essential

[Comment: Short-term predictions are much more likely to be valid than are long-term predictions because they are less subject to influences other than the characteristics measured.]

E7.4.3. If a test is recommended for long-term predictions, but comparisons with concurrent criteria only are presented, the manual should emphasize that the validity of long-term predictions is undetermined. Essential

E7.4.4. The amount and kind of any experience or training received by the subjects between the time of testing and the time of criterion measurement should be stated. Very Desirable

E7.4.5. When validity for predicting grades in a course is reported, reasonably clear information should be provided regarding the types of performance required in the course, the nature of the instructional method, and the way in which performance is measured. If the test was administered after the course was started, this fact should be made clear. Very Desirable

E8. Any statistical analysis of criterion-related validity should be reported in the manual in a form that enables the reader to determine how much confidence is to be placed in judgments or predictions regarding the individual. Essential

E8.1. A report of criterion-related validity should give full information about the statistical analysis and should ordinarily include, in addition to such basic descriptive statistics as means and standard deviations, one or more of the following: (*a*) one or more correlation coefficients of a familiar type, (*b*) descriptions of the efficiency with which the test separates criterion groups, (*c*) expectancy tables, or (*d*) charts that graphically illustrate the relationship between test and criterion. Essential

[Comment: Full information includes data on the reliability, the strength, and the nature of the relationship. In correlational terms, this would imply information about the statistical significance and magnitude of the correlation coefficient and about the regression equation.

Reports solely of differences between group means give inadequate information regarding validity; if variance is large, classification may be inaccurate even if means differ considerably. The strength of the relationship may be indicated by describing the amount of misclassification or of overlapping. Expectancy tables may provide information about the nature of the predictions.

In general, since manuals and research reports are often directed to test users who have limited statistical knowledge, every effort should be made to communicate validity information clearly.]

E8.1.1. Errors of prediction should be estimated and reported; a validity coefficient should be supplemented with reports of the regression slope and intercept and of the standard error of estimate. Very Desirable

[Comment: The required information could be presented in an expectancy table showing the range of possible criterion values for each of several points on the score range. The standard error of estimate at different points along the score range is often helpful.

For a dichotomous criterion, this objective might be achieved by indicating the proportion of hits, misses, and false inclusions at various cutting scores.]

E8.1.2. For some users, analysis of test variance according to the following sources is appropriate: variance relevant to the criterion, variance explained as form-to-form or trial-to-trial inconsistency, and a reliable but irrelevant remainder. Very Desirable

[Comment: Such an analysis is more complete and less subject to misinterpretation than a correlation coefficient, including even a "corrected" validity coefficient, or a comparison of group means.]

E8.1.3. The method of statistical analysis should be chosen with due consideration of the characteristics of the data and of the assumptions of the method. Essential

[Comment: Data may often depart from the assumed characteristics with little ill effect. Some violation of

to communicate validity information clearly.]

E8.1.1. Errors of prediction should be estimated and reported; a validity coefficient should be supplemented with reports of the regression slope and intercept and of the standard error of estimate. Very Desirable

[Comment: The required information could be presented in an expectancy table showing the range of possible criterion values for each of several points on the score range. The standard error of estimate at different points along the score range is often helpful.

For a dichotomous criterion, this objective might be achieved by indicating the proportion of hits, misses, and false inclusions at various cutting scores.]

E8.1.2. For some users, analysis of test variance according to the following sources is appropriate: variance relevant to the criterion, variance explained as form-to-form or trial-to-trial inconsistency, and a reliable but irrelevant remainder. Very Desirable

[Comment: Such an analysis is more complete and less subject to misinterpretation than a correlation coefficient, including even a "corrected" validity coefficient, or a comparison of group means.]

E8.1.3. The method of statistical analysis should be chosen with due consideration of the characteristics of the data and of the assumptions of the method. Essential

[Comment: Data may often depart from the assumed characteristics with little ill effect. Some violation of assumptions may, however, be seriously misleading. For example, the use of predictions based on the assumption of a normal bivariate correlation surface may seriously overestimate the mean performance of high-scoring candidates if the data are markedly heteroscedastic (as in triangular scatter distributions). In such cases, a method of analysis not based on assumptions about the bivariate distribution would present a more accurate statement of validity.]

E8.2. If validity coefficients are corrected for errors of measurement *in the criterion*, the computation of the reliability coefficient of the criterion should be explained, and both corrected and uncorrected coefficients should be reported. Essential

[Comment: Coefficients corrected for errors of measurement in the *test* are not estimates of the criterion-related validity for the existing test and should not be reported. Corrections for attenuation are very much open to misinterpretation, especially if based on obtained correlation coefficients that are very low or from a small or otherwise inappropriate sample; if misinterpreted, they give an unjustifiably favorable impression of the validity of the test scores. The hazard is illustrated in the manual for an adjustment inventory. The author reported correlation coefficients between inventory scores and criterion ratings; also reported were estimated coefficients between "true" inventory and criterion scores. He then commented that the augmented correlation coefficients "are as high as those often secured between college aptitude tests and college grades." The comparison is improper, in part, because the test author compared augmented coefficients with uncorrected coefficients for ability tests.]

E8.2.1. Where correlation coefficients are corrected for attenuation or restricted range, full information relevant to the correction should be presented. If such corrections are made, significance tests should be made with the uncorrected correlation coefficients. Essential

[Comment: Corrections should be applied only to obtained coefficients. It is ordinarily unwise to make sequential corrections, as in applying a correction for attenuation to a coefficient already corrected for restriction of range. Chains of corrections may be useful in considering possible further research, but their results should not be seriously reported as estimates of population correlation coefficients.]

E8.3. If validity is demonstrated by comparing groups that differ on the criterion, the manual should report whether and by how much the groups differ on other available variables that are relevant. Very Desirable

[Comment: Since groups that differ on a criterion may also differ in other respects, the test may be discriminating on a quality other than that intended. Types of mental disorders, for instance, are associated with age, education, and length of time in the hospital. Confounding of this sort should be taken into account when the usefulness of a test for diagnosis is appraised.]

E8.3.1. If a test is suggested for the differential diagnosis of patients, the

manual should include evidence of the test's ability to place individuals in diagnostic groups rather than merely to separate diagnosed abnormal cases from the normal population. Essential

[Comment: When a test is recommended for the purpose of assigning patients to discrete categories, such statistics as contingency coefficients, phi coefficients, or discriminant functions should be supplemented by a table of misclassification rates giving, for example, the proportion of patients falsely included in a category or falsely excluded from it. Such proportions should be compared with base rates, that is, the proportions of correct classifications made possible by a mere knowledge of the sizes of the categories.]

E8.3.2. If validity is demonstrated by comparing groups that differ on the criterion (e.g., where one group is identified as a high-performance group and another as a low-performance group), all cases should be assigned to one or the other of the groups. Very Desirable

[Comment: The most reliable statistics are obtained if all cases are used; validity coefficients derived from extreme groups may be misleading. In some situations, analyses using extreme groups may be useful for identifying predictors, but generally the validity reported for any given predictor should be based on all cases. If the use of extreme groups is deemed necessary or appropriate to a particular study, appropriate estimates of correlation should be used. The typical product-moment and biserial estimates are *not* appropriate in this situation.]

E8.4. When information other than the test scores is known to have an appreciable degree of criterion-related validity and is ordinarily available to the prospective test user, the user should consider both the validity of the other information and the resulting multiple correlation when the new test information is combined with it. Essential

[Comment: Whether a test should be used for prediction and classification when other information is readily available sometimes depends not on the validity of the test but on its "incremental validity," that is, what it adds to the soundness of the judgment that would otherwise be made.

For a questionnaire intended to predict marital success, delinquency, and similar behavioral variables, the investigator should find out how much the questionnaire enhances prediction over that provided by base rates developed from demographic variables such as socioeconomic status.]

E8.5. Where more than one test is to be used, validity information should report the validity of the combination actually used. Where composite scores are developed, the basis for weighting (e.g., multiple regression equations) should be given. Essential

[Comment: In one organization, a composite was developed and validated by multiple regression in which the optimal weighting of one test was negative. Nevertheless, the organization added unweighted scores to form a different composite for use in making decisions. The multiple correlation coefficient did not, therefore, describe the validity of the test battery as it was

E8.4. When information other than the test scores is known to have an appreciable degree of criterion-related validity and is ordinarily available to the prospective test user, the user should consider both the validity of the other information and the resulting multiple correlation when the new test information is combined with it. Essential

[Comment: Whether a test should be used for prediction and classification when other information is readily available sometimes depends not on the validity of the test but on its "incremental validity," that is, what it adds to the soundness of the judgment that would otherwise be made.

For a questionnaire intended to predict marital success, delinquency, and similar behavioral variables, the investigator should find out how much the questionnaire enhances prediction over that provided by base rates developed from demographic variables such as socioeconomic status.]

E8.5. Where more than one test is to be used, validity information should report the validity of the combination actually used. Where composite scores are developed, the basis for weighting (e.g., multiple regression equations) should be given. Essential

[Comment: In one organization, a composite was developed and validated by multiple regression in which the optimal weighting of one test was negative. Nevertheless, the organization added unweighted scores to form a different composite for use in making decisions. The multiple correlation coefficient did not, therefore, describe the validity of the test battery as it was actually used. Where a given method of combination is to be used, that method should be validated.

When multiple regression is used and one predictor in a battery is evaluated, the beta weight is a better index of its contribution to the validity of the test in that combination than is its original validity coefficient.]

E9. A test user should investigate the possibility of bias in tests or in test items. Wherever possible, there should be an investigation of possible differences in criterion-related validity for ethnic, sex, or other subsamples that can be identified when the test is given. The manual or research report should give the results for each subsample separately or report that no differences were found. Essential

[Comment: For many uses, regulations published pursuant to civil rights legislation require that validity studies be performed separately on samples differing in national origin, race, sex, or religious affiliation, when technically feasible.

The concept of fairness may involve other sources of inappropriate discrimination. For example, placing a hand-dexterity test on a low table may unfairly bias the test against tall people. The test user should try to identify potentially unfair influences on test scores in his situation. Variables which may contribute inappropriate variance may be used for subgrouping in investigation of fairness.

However, caution must be exercised in evaluating the possibility of bias. A simple difference in group means does *not* by itself identify an unfair test, although it should stimulate research to explore the question of fairness. Evidence of differential validity is developed by comparing, for example, correlation coefficients, regression equations, and means and variances for each variable.

The proper statistical test for such a difference is, for any parameter, the test of the hypothesis of no true difference between the groups, for example, a test of no difference between correlation coefficients, slopes, or intercepts. Some investigators have attempted to examine such differences by comparing in each subgroup independently the validity statistic (e.g., the correlation coefficient) to a postulated true value of zero. This is not a proper procedure; it does not answer the question at issue of *differences* in the characteristics of validity. It is impossible to demonstrate such differences by showing that one correlation coefficient, for example, is significantly different from zero while the other is not.

Users should routinely investigate differences in validity when it is technically feasible to do so, that is, when Ns are sufficient for reliable comparisons and when criteria are reasonably valid in each group. Users should be aware, however, that a too-hasty acceptance of bias or of differential validity, if used in decision making, may be as likely to produce unfair test use as is failure to consider the possibility.

For example, to avoid unfairness in test use for blacks, an employer may investigate the possibility of differential validity and find not only differences in means between black and white applicants but also differences in intercepts of the regression. Some definitions of fairness require that

predictions for applicants in either group should be based on the regression line developed for his own group. If the differences in intercepts are statistical artifacts (due, for example, to unreliability), the result might be considered unfair to blacks (if they have the lower regression line) since their performance might be systematically under predicted. The effect can, of course, work both ways depending on the direction of differences in regression.

It is important to recognize that there are different definitions of fairness, and whether a given procedure is or is not fair may depend upon the definition accepted. Moreover, there are statistical and psychometric uncertainties about some of the sources of apparent differences in validity or regression. Unless a difference is observed on samples of substantial size, and unless there is a reasonably sound psychological or sociological theory upon which to explain an observed difference, the difference should be viewed with caution.

Bias is not necessarily detected by criterion-related validity alone; cf. E12.12).

E10. When a scoring key, the selection of items, or the weighting of tests is based on one sample, the manual should report validity coefficients based on data obtained from one or more independent cross-validation samples. Validity statements should not be based on the original sample. Essential

E10.1. If the user recommends certain regression weights for combining scores on a test or for combining the test with other variables, the statement of the validity of the composite should be based on a crossvalidation sample. Essential

[Comment: Cross-validation is particularly necessary when the number of predictors entering the study (not the final equation) is greater than 4 or 5 and when the sample size is less than 200.]

E10.1.1. When the scoring of tests in a battery is based on regression coefficients, negative scoring weights should be used only if they have been verified by cross-validation in large samples and if their use will not be invalid (and thus unfair) to one or more subgroups in the population to be tested. Essential

E10.2. If it is proposed that decisions be based on a complex nonlinear combination of scores, it should be shown that this combination has greater validity than a simpler linear combination, that the equation can be logically explained, and that the procedures for combining scores have been cross validated. Essential

[Comment: The use of "moderator variables," for example, is to be recommended only where a moderator is shown to produce a clear improvement in validity in a cross-validation sample. Similarly, when it is proposed that some pattern of scores (e.g., high standing in scores on both Variables 2 and 5) is an indicator of success, it is necessary to show that the proportion of successful persons in the group so identified is higher than would be expected from the regression of

e test with other variables, the state-ent of the validity of the composite ou be based on a crossvalidation mple. Essential

[Comment: Cross-validation is par-cula y necessary when the number of edictors entering the study (not the al equation) is greater than 4 or 5 d en the sample size is less than 00.]

E10.1.1. When the scoring of tests a attery is based on regression efficients, negative scoring weights ould be used only if they have been erif by cross-validation in large mp and if their use will not be valid (and thus unfair) to one or more bgroups in the population to be ste Essential

E10.2. If it is proposed that ecis s be based on a complex onli ar combination of scores, it ould be shown that this combination as greater validity than a simpler nea ombination, that the equation n b ogically explained, and that the rocedures for combining scores have een cross validated. Essential

[Comment: The use of "moderator riables," for example, is to be ecommended only where a moderator wn to produce a clear npr ment in validity in a cross-lidation sample. Similarly, when it is oposed that some pattern of scores g., gh standing in scores on both ari es 2 and 5) is an indicator of ccess, it is necessary to show that the opo ion of successful persons in the ou identified is higher than would expected from the regression of

frequency of success on a linear combination of Variables 2 and 5.]

E11. To the extent feasible, a test user who intends to continue employing a test over a long period of time should develop procedures for gathering data for continued research. Desirable

[Comment: Validity data may become obsolete. The relationship between test performance and criterion performance may be influenced by many factors, such as changes in populations, recruiting sources, the economy, organizational characteristics, processes, or tasks. Moreover, validity studies are often based on relatively few cases. A plan for the systematic collection of further data after the test has been placed into operational use may be useful both for the development of a more reliable data base and for information on changes in the trends of relationships over time.

Operational use may, however, result in severe restriction of range. Continuing research may be less necessary if the original data are based on a relatively large sample, if the bases for generalizing validity are well established, and if evidence shows a relatively slight rate of change in variables likely to limit the generizability of validity information. When these favorable conditions do not exist, it may be possible to plan for small replications from time to time rather than for a continuous program of research.]

Content Validity

E12. If test performance is to be interpreted as a representative sample of performance in a universe of situations, the test manual should give a clear definition of the universe represented and describe the procedures followed in the sampling from it. Essential

[Comment: The definition of the universe of tasks represented by the test scores should include the identification of that part of the content universe represented by each item. The definition should be operational rather than theoretical, containing specifications regarding classes of stimuli, tasks to be performed and observations to be scored. The definition should not involve assumptions regarding the psychological processes employed since these would be matters of construct rather than of content validity.]

E12.1. When experts have been asked to judge whether items are an appropriate sample of a universe or are correctly scored, the manual should describe the relevant professional experience and qualification of the experts and the directions under which they made their judgments. Very Desirable

E12.1.1. When items are selected by experts, the extent of agreement among judges should be reported. Desirable

E12.1.2. Test content should be examined for possible bias. Essential

[Comment: Bias may exist where items do not represent comparable tasks and therefore do not sample a common performance domain for the various subgroups (cf. B1.3). One may investigate such bias in terms of carefully developed expert judgments; studies of the attitudes or

interpretations of items in different subgroups might also present useful information (although care must be taken to assure that the investigation is clearly directed to an analysis of content in relation to an adequately defined performance domain). The judgment of bias may itself be biased; the principle here is that, when it is possible, such judgments should be supported by data.]

E12.2. In achievement tests of educational outcomes, the manual should report the classification system used for selecting items. Desirable

E12.2.1. When an achievement test has been prepared according to a two-way content-by-process outline, that outline should be presented in the manual, with a list of the items identified with each cell of the outline. Very Desirable

E12.3. Any statement in the manual of the relation of items to a course of study (or other source of content) should mention the date when the course of study was prepared. Essential

[Comment: In achievement testing, it is frequently the practice to identify significant topics for items by a careful sampling from textbooks. Textbooks and courses of study change, however, and the test that was once an excellent sample becomes out of date. The manual might therefore report such information as the range and median of copyright dates of the textbooks examined, or the date at which the experts judged the items to be representative.

One checklist concerns problems common to students. The manual for this checklist properly reports the date when the list was assembled. From time to time, it will be necessary to determine whether student problems have changed and, if so, to change the test accordingly.

It should be recognized that this standard implies that definitions of a content universe are subject to change as jobs, society, or curricula change.]

E12.4. When a test is represented as having content validity for a job or class of jobs, the evidence of validity should include a complete description of job duties, including relative frequency, importance, and skill level of such duties. Essential

Construct Validity

E13. If the author proposes to interpret scores on a test as measuring a theoretical variable (ability, trait, or attitude), his proposed interpretation should be fully stated. His theoretical construct should be distinguished from interpretations arising on the basis of other theories. Essential

[Comment: For example, if a test is intended to measure the construct of anxiety, the test author should distinguish his formulation of the construct from other possible meanings of the term and should relate his concept to measures of anxiety discussed in the literature.

The description of a construct may be as simple as the identification of "creativity" with "making many original contributions." Even this definition provides some basis for judging whether various pieces of

mon to students. The manual for checklist properly reports the date n th ist was assembled. From time me, will be necessary to determine ther student problems have ged nd, if so, to change the test rdi y.

should be recognized that this dard implies that definitions of a ent niverse are subject to change bs, ciety, or curricula change.]

12.4. When a test is represented avi content validity for a job or s o obs, the evidence of validity uld include a complete description job duties, including relative uen , importance, and skill level of duties. Essential

strt Validity

13. If the author proposes to rpret scores on a test as measuring a oret l variable (ability, trait, or tud his proposed interpretation uld be fully stated. His theoretical struct should be distinguished from rpr tions arising on the basis of er t ries. Essential

Comment: For example, if a test is end to measure the construct of iet the test author should tinguish his formulation of the struct from other possible meanings the term and should relate his ncep to measures of anxiety cussed in the literature.

The scription of a construct may be si le as the identification of eativity" with "making many ginal contributions." Even this fini n provides some basis for dgi whether various pieces of empirical evidence support the proposed interpretation. Ordinarily, however, the test author will have a more elaborate conception. He may wish to rule out such originality as derives only from a large and varied store of information. He may propose explicitly to identify the creative person as one who produces numerous ideas, whether of high or low quality. He may propose to distinguish the ability to criticize ideas from the ability to be "creative." He may go on to hypothesize that the person who shows originality in identifying or describing pictures will also have unconventional preferences in food and clothing. All such characterizations or hypotheses are part of the author's concept of "what the test measures" and are needed in designing and in drawing conclusions from empirical investigations of the psychological interpretation of the construct.]

E13.1. The manual should indicate the extent to which the proposed interpretation has been substantiated and should summarize investigations of the hypotheses derived from the theory. Essential

E13.1.1. Each study investigating a theoretical inference regarding the test should be summarized in a way that covers both the operational procedures of the study and the implications of the results for the theory. Very Desirable

E13.1.2. The manual should report correlations between the test and other relevant tests for which interpretations are relatively clear. Very Desirable

E13.2. The manual should report evidence about the extent to which constructs other than those proposed by the author account for variance in scores on the test. Very Desirable

[Comment: Although it is unreasonable to require a test author to anticipate or to include every counterinterpretation in a test manual, he ought to present data relevant to those counterhypotheses most likely to account for variance in the test scores.]

E13.2.1. The manual for any specialized test or inventory used in educational selection and guidance should report the correlation of scores derived from it with well-established measures of verbal and quantitative ability in an appropriately representative population. Very Desirable

[Comment: Verbal and quantitative abilities are specified here because their importance in educational performance is recognized, because they often account for much of total test variance, and because numerous tests of these abilities are already available. To be of practical value, a new test designed to measure other constructs (e.g., spatial abilities) must not closely duplicate the measurement of verbal and quantitative ability.]

E13.2.2. If a test has been included in factorial studies that indicate the proportion of the test variance attributable to widely known reference factors, such information should be presented in the manual. Desirable

E13.2.3. For inventories such as personality, interest, or attitude measures, evidence should be presented of the extent to which scores are susceptible to an attempt by the examinee to present a socially desirable,

conforming, or false picture of himself, or to which the scores may reflect other response sets or styles. Such response patterns should be studied for identifiable subgroups rather than for a more general sample. Very Desirable

[Comment: Correlational or experimental studies might be reported. Appropriate evidence of acquiescence might, for example, be the proportion of the total test variance in the number of yes responses to the test, or by the correlation of the test scores with one or more independent measures of the acquiescence tendency, or by experimental procedures designed to induce acquiescence.]

E13.2.4. If a test given with a time limit is to be interpreted as measuring a hypothetical psychological attribute not specifically related to speed, evidence should be presented in the manual concerning the effect of speed on the test scores and on their correlation with other variables. Essential

[Comment: The most complete evidence of the effect of speed would be the comparison of scores on one form, using the usual time limit, with scores on another form having unlimited time. The correlation of scores at the end of the usual time with scores obtained with extra time on the same trial is of limited meaning because the two scores are not independent. Less complete evidence would consist of data on the percentage of examinees who attempt the last item or some item very near the end of the test. If the percentage is below 90, a more penetrating study is needed to show that individual differences on the test do not reflect speed to any great extent.]

E13.2.5. Where differences in test-taking strategies that might influence the interpretation of scores are associated with identifiable subgroup characteristics, this information should be clearly presented or its absence clearly noted. Very Desirable

E13.2.6. Where a low correlation or small difference between groups is advanced as evidence *against* some counterinterpretation, the manual should report the confidence interval for the parameter. The manual should also correct for or discuss any errors of measurement that may have lowered the apparent relationship. Desirable

F. Reliability and Measurement Error

Reliability refers to the degree to which the results of testing are attributable to systematic sources of variance. Classical methods of estimating reliability coefficients call for correlating at least two sets of similar measurements.

One method of obtaining the two sets of measurements is by retesting with the identical test. Aside from practical limitations, theoretically, retesting is not ordinarily a desirable method of estimating reliability because the examinee may remember his or her responses to items from one testing to the next. Hence, memory becomes a systematic source of variance and the correlation of the two sets of scores may be higher than the correlation of two sets of scores based on two different but parallel sets of items drawn from the population of items in the same way.

If we want to eliminate memory as a systematic source of variance and to

DS

E13.[illegible]. Where differences in testing strategies that might influence interpretation of scores are associated with identifiable subgroup characteristics, this information should be clearly presented or its absence clearly noted. Very Desirable

E13.[illegible]. Where a low correlation or small difference between groups is analyzed as evidence *against* some interpretation, the manual should report the confidence interval for the parameter. The manual should also correct for or discuss any errors of measurement that may have lowered the apparent relationship. Desirable

F. Reliability and Measurement Error

Reliability refers to the degree to which the results of testing are attributable to systematic sources of variance. Classical methods of estimating reliability coefficients call for correlating at least two sets of similar measurements.

One method of obtaining the two sets of measurements is by retesting with the identical test. Aside from practical limitations, theoretically, retesting is not ordinarily a desirable method of estimating reliability because the examinee may remember his or her responses to items from one testing to the next. Hence, memory becomes a systematic source of variance and the correlation of the two sets of scores may be higher than the correlation of two sets of scores based on two different but parallel sets of items drawn from the population of items in the same way.

If we want to eliminate memory as a systematic source of variance and to include the effects of item sampling and response variation over time as sources of variance, we may use two sets of items developed or selected according to the same specifications. These are called parallel forms of the test.

If the effect of content sampling *alone* is sought without the effects of memory or response variability over time, or if it is not practical to administer two parallel forms with separate time limits, reliability can be estimated from a single administration of an unspeeded test. The test may be divided into two sets of items of equal, or approximately equal, length that are judged by competent authorities to sample as nearly as possible the same functions. Any items based on the same source of data (such as a reading passage) must be assigned to the same set. Then the correlation between scores on the two parallel halves is a matched-half coefficient from which an estimate of the parallel-forms reliability coefficient for the total test may be obtained by a procedure that does not assume that the numbers of items or the variances of the two sets are exactly equal.

Estimates of reliability from a single administration may also be obtained by analysis-of-variance procedures. Such estimates will be spuriously high if the test is speeded or if the items are not independent of each other. On the other hand, for unspeeded tests, such estimates will tend to be lower than matched-half coefficients because they constitute, given certain assumptions, the mean of coefficients obtained by correlating scores on all possible pairs of halves of the test.

From the preceding discussion, it is clear that *different methods of estimating reliability take account of different sources of error.* Thus, from one testing to the other, the result is affected not only by random response variability and changes in subjects over time but also by differences in administration (especially if different persons administer the test on the two occasions). Reliability coefficients based on a single administration of a test exclude response variability over time; these effects on scores do not appear as errors of measurement. Hence, "reliability coefficient" is a generic term. It can be based on various types of evidence; each type of evidence suggests a different meaning. It is essential that any method used to estimate reliability be clearly described.

The estimation of clearly labeled components of score variance is the most informative outcome of a reliability study, both for the test developer wishing to improve the reliability of his instrument and for the user desiring to interpret test scores with maximum understanding. The analysis of score variance calls for the use of an appropriate experimental design. There are many different multivariate designs that can be used in reliability studies; the choice of design for studying a particular test is determined by its intended interpretation and by practical limitations.

It is recommended that test authors describe the meanings of any coefficients they report as accurately and precisely as possible. It is informative to say, for example, "This coefficient indicates the stability of measurement of equivalent scores based on parallel forms of the test

administered 7 days apart, without intervening practice or instruction." Although lengthy, such a description is reasonably free from ambiguity.

Reliability coefficients have limited practical value for test users. The standard error of measurement ordinarily is more useful; it has great stability across populations since it is relatively independent of range of talent, and it may be used to identify limits that have a defined probability of including the true score. Test users may use reliability coefficients in comparing tests, but they use standard errors of measurement in interpreting test scores. Information in a test manual about a standard error of measurement may often be more important than information about a reliability coefficient.

General Principles

F1. The test manual or research report should present evidence of reliability, including estimates of the standard error of measurement, that permits the reader to judge whether scores are sufficiently dependable for the intended uses of the test. If any of the necessary evidence has not been collected, the absence of such information should be noted. Essential

[Comment: It is most helpful to the user when several types of reliability estimates are reported. Reports of standard errors of measurement in different groups are also helpful.]

F1.1. The test manual should furnish, insofar as feasible, a quantitative analysis of the total inconsistency of measurement into its major identifiable components; namely, inconsistency in responses of the subject; inconsistency or hetereogeneity within the sample of test content (such as the stimulus items, questions, and situations); inconsistencies in administration of the test; inconsistency among scorers, raters, or units of apparatus; and mechanical errors of scoring. Desirable

[Comment: In general, the desired analysis will not be feasible unless scores are expressed in quantitative, as distinguished from categorical or nonparametric, terms and the design of data collection includes the necessary controls.

With group tests of school achievement, the principal sources of error to be evaluated usually include: (*a*) inconsistency of test content; (*b*) inconsistencies in test administration; and (*c*) inconsistency in responses of the examinee over time, that is, instability. The collection of data should be designed to permit evaluation of these three factors. Fluctuation or inconsistency in the responses of the subject may be an important variable by itself; it is often a major source of random error to be evaluated. Inconsistency among scorers or raters should also be evaluated.]

F1.2. Standard errors of measurement and reliability coefficients should be provided for every score, subscore, or combination of scores (such as a sum, difference, or quotient) that is recommended by the test manual (either explicitly or implicitly) for other than merely tentative or pilot use. Essential

F1.3. For instruments that yield a profile having a low reliability of

ajor entifiable components; namely, ons ency in responses of the oject; inconsistency or hetereogeneity hin the sample of test content (such the timulus items, questions, and uati s); inconsistencies in admin-ration of the test; inconsistency ong scorers, raters, or units of appa-tus; d mechanical errors of scoring. sirable

Comment: In general, the desired alys will not be feasible unless res e expressed in quantitative, as tinguished from categorical or npar metric, terms and the design of a c ection includes the necessary trols.

With group tests of school iev ent, the principal sources of or be evaluated usually include: inconsistency of test content; inconsistencies in test adminis-tion and (c) inconsistency in re-onse of the examinee over time, that instability. The collection of data uld designed to permit evaluation the three factors. Fluctuation or onsistency in the responses of the oject may be an important variable itse it is often a major source of dom error to be evaluated. Inconsis-cy among scorers or raters should o be evaluated.]

F1.2 Standard errors of measure-nt and reliability coefficients should provided for every score, subscore, or nbi ion of scores (such as a sum, fere e, or quotient) that is recom-nded by the test manual (either ex-citly or implicitly) for other than rely ntative or pilot use. Essential

F1.3. For instruments that yield a file having a low reliability of differences between scores, the manual should explicitly caution the user against interpretation of such differences, except as a source of tentative information requiring external verification. Essential

F1.4. The manual should state the minimum difference between two scores ordinarily required for statistical significance at a designated level. Very Desirable

[Comment: A nomograph or table for determining the significance of any given score difference would be a very useful addition to a test manual. "Change" or "growth" scores require careful attention to Standards F1.3 and F1.4.]

F2. The procedures and samples used to determine reliability coefficients or standard errors of measurement should be described sufficiently to permit a user to judge the applicability of the data reported to the individuals or groups with which he is concerned. Essential

[Comment: The mean and variance of the sample and information about its composition should be provided. Reliability data should be obtained from "natural" groups such as examinees of a single age or grade level. Estimates of the reliability of a test to be used in selecting employees should be based on scores of applicants for positions rather than scores obtained by testing college students or workers already employed.

If a test claims to be appropriate for groups from the fourth grade through graduate school, the manual should provide reliability data for each grade or age level.]

F2.1. Any identifying characteristics of the sample that may be related to consistency of performance on the test should be described in the test manual. Essential

[Comment: Demographic information, such as distributions of the subjects with respect to age, sex, socioeconomic level, intellectual level, locale, employment status or history, and minority group membership should be given in the test manual. For standardized tests, the samples used to compute reliability coefficients and standard errors of measurement should be drawn at random from the norms groups.]

F2.2. If reliability coefficients are corrected for restriction of range, both the uncorrected and the corrected coefficients should be reported in the test manual together with the standard deviations of the group actually tested and of the group to which the corrected coefficients are applicable. Essential

[Comment: When variances differ and there seem to be other justifications for such a correction, the superiority of the standard error of measurement should be noted; it is largely unaffected by differences in variance.]

F2.3. When a test is recommended or ordinarily employed in homogeneous subsamples, the reliability and standard error of measurement should be independently investigated within each subsample and reported in the test manual. Essential

[Comment: The mechanical reasoning section of a well-known aptitude test yields scores that have significantly different reliability coefficients for boys

and for girls. The manual reports the reliability coefficients for each sex in each grade.]

F2.3.1. At least one estimate of the standard error of measurement should be provided in the manual for every group for which reliability data are given. Essential

[Comment: When it is specifically recommended that scores be transformed to a particular metric, the standard errors should be presented in that metric.]

F2.3.2. The test manual should report the standard errors of measurement at different score levels. Desirable

[Comment: The manual for one test of college aptitude reports standard errors of measurement for three score levels: the mean, one standard deviation above the mean, and one standard deviation below the mean. Since more important changes in the standard error of measurement are associated with extreme scores, it might be better to use more widely separated score levels if the number of cases available justifies this action.

F2.4. Item statistics (such as difficulty or discrimination indices, etc.) should be presented in at least summary form in a test manual. Desirable

F3. Reports of reliability studies should ordinarily be expressed in the test manual in terms of variances of error components, standard errors of measurement, or product-moment reliability coefficients. Unfamiliar expressions of data should be clearly described, with references to their development. Essential

[Comment: Test authors and publishers should avoid unconventional statistics unless conventional statistics are inappropriate. If unusual statistical analyses are presented, explanations should minimize the likelihood of misinterpretation.]

Comparability of Forms

F4. If two or more forms of a test are published for use with the same examinees, information on means, variances, and characteristics of items in the forms should be reported in the test manual along with the coefficients of correlation among their scores. If necessary evidence is not provided, the test manual should warn the reader against assuming equivalence of scores. Essential

[Comment: Information to be examined would include a summary of item statistics for each form, such as a frequency distribution of item difficulties and of indices of item discrimination. Content analyses of each of the forms should be presented. Thus, both frequency distributions of item statistics and a tabulation of items by categories of subject-matter content and of behavioral or instructional objectives should be furnished.

The forms should represent different samples of items within each category of content. Insofar as one's concern is for error arising from sampling a content universe, the forms to be compared should have been developed from a common universe according to an appropriate plan. An artificially

close similarity between forms will result from item-by-item matching or by creating a second form merely by rephrasing items on a first form. A reliability coefficient based on forms created in this way will be spuriously high because it does not properly take into account sampling error in drawing items from the universe of items.]

Internal Consistency

F5. Evidence of internal consistency should be reported for any unspeeded test. Very Desirable

[Comment: Internal consistency is important if items are viewed as a sample from a relatively homogeneous universe, as in a test of addition with integers, a test of general high school vocabulary, or a test presumed to measure introversion. Nevertheless, estimates of internal consistency should not be regarded as a substitute for other measures.]

F5.1. Estimates of internal consistency should be determined by matched-half or random-half methods or by analysis of variance procedures, if these can properly be used with the data. Any additional measure of internal consistency that the author wishes to report should be carefully explained in the test manual. Very Desirable

[Comment: Matched-half coefficients reflect expert judgment and tend to be higher in value than random-half coefficients. Analysis of variance procedures tend to yield lower values than matched-half procedures. In unusual circumstances, special coefficients may provide useful information; if used, such coefficients should be described so the reader will be able to understand them in relation to more conventional estimates.]

F5.2. Internal reliability estimates should not be obtained for highly speeded tests. Essential

F5.3. When a test consists of separately scored parts or sections, the correlation between the parts or sections should be reported in the test manual along with relevant reliability estimates, relevant means, and relevant standard deviations. Very Desirable

F5.3.1. If a test manual reports the correlation between a subtest and a total score, it should call attention to the fact that the coefficient is spuriously high because it is based partly on the perfect correspondence of identical errors of measurement in the subtest and in the total score. Essential

F5.4. If several questions within a test are experimentally linked so that the reaction to one question influences the reaction to another, the entire group of questions should be assigned to one of the two halves of the test when random-half or matched-half procedures are used. Very Desirable

[Comment: In a reading test, several questions about the same paragraph are ordinarily experimentally dependent. All of these questions should be placed in the same half test in using the split-half method. The fact that the test halves do not have exactly equal numbers of items need not be troublesome if an appropriate step-up procedure is used.]

Comparisons Over Time

F6. The test manual should indicate to what extent test scores are stable, that is, how nearly constant the scores are likely to be if a parallel form of a test is administered after time has elapsed. The manual should also describe the effect of any such variation on the usefulness of the test. The time interval to be considered depends on the nature of the test and on what interpretation of the test scores is recommended. Essential

[Comment: For many purposes, reliability coefficients and standard errors of measurement should be based on parallel-forms procedures, with a period of perhaps 2 to 4 weeks elapsing between the administration of two parallel forms. In some situations, when test scores are obtained for changing characteristics of individuals, reliability coefficients or standard errors of measurement based on the administration of parallel forms on successive days or weeks may be desirable. A reading-readiness test used only for initial tentative assignment of first-grade pupils to instructional groups is an example. In experiments on the effects of drugs, it may be desirable to measure changes in two sets of test scores obtained before and after a time lapse of only a few minutes.

It seems reasonable to require an assessment of stability for projective techniques and other devices for assessing personality dynamics, even though it is recognized in some instances that low stability of scores over a substantial period may reflect true trait fluctuation. Clinical practice rarely presumes that the inferences from projective tests are to be applied on the very day the test is given. Realistically, one must recognize that pragmatic decisions are being made from test data which are meaningful only in terms of at least days, and usually weeks or months of therapy. If scores on a certain test are found to be highly unstable from day to day, this evidence casts doubt upon the utility of the test for most purposes, even if some fluctuation might be explained by the hypothesis of trait inconstancy. An investigator may be concerned with a psychological characteristic or educational effect which changes rapidly over a short period of time. In this instance it is important not to confuse the inconstancy of the trait with the instability of the measuring instrument.]

F6.1. Determination of the stability of scores by repeated testing should make use of parallel forms of the test to minimize recall of specific answers, especially if the time interval is short. Very desirable

F6.2. The report in a test manual of a study of consistency of scores over time should state what period of time elapsed between tests and should give the mean and standard deviation of scores at each testing as well as the correlation coefficient. Essential

F6.3. If it is reasonable to expect scores on a test to change significantly over some time interval in response to developmental or educational influences, the manual should call the test user's attention to this possibility and advise care in the use of old scores. Very Desirable

[Comment: Since some schools

administer
interest
years, th
report
means a
tests ad
apart, a
data the u
records
passage of

F6.3.1.

n the very day the test is given. Realistically, one must recognize that pragmatic decisions are being made from test data which are meaningful only in terms of at least days, and usually weeks or months of therapy. If scores on a certain test are found to be highly unstable from day to day, this widely casts doubt upon the utility of the test for most purposes, even if some fluctuation might be explained by the hypothesis of trait inconstancy. An investigator may be concerned with a psychological characteristic or educational effect which changes rapidly over a short period of time. In this instance it is important not to confuse the inconstancy of the trait with the instability of the measuring instrument.]

F6.1. Determination of the stability of scores by repeated testing should make use of parallel forms of the test to minimize recall of specific answers, especially if the time interval is short. Very desirable

F6.2. The report in a test manual of a study of consistency of scores over time should state what period of time elapsed between tests and should give the mean and standard deviation of scores on each testing as well as the correlation coefficient. Essential

F6.3. If it is reasonable to expect scores on a test to change significantly over some time interval in response to developmental or educational influences, the manual should call the test user's attention to this possibility and advise caution in the use of old scores. Very desirable

Comment: Since some schools

administer aptitude, achievement, or interest tests only at intervals of 2 or 3 years, the manual for such tests should report correlations and changes in means and standard deviations between tests administered 1 year apart, 2 years apart, and 3 years apart. From these data the user can learn how rapidly test records become obsolete with the passage of time.]

F6.3.1. In reporting on stability, the test manual should describe relevant experience, education, or treatment intervening between administrations of the test, if known. Desirable

F6.4. Where a test is to be used to compare groups rather than individuals, standard errors and standards errors of measurement of group means and related statistics should be presented. Essential

STANDARDS FOR THE USE OF TESTS

There are many kinds of test use. As one example, test scores are used for decisions to select or to reject applicants for jobs, schools, or other opportunities. In such use, the test score is a basis for a prediction, one that is either explicit or strongly implied. The test score is used to estimate or predict a likely level of performance on some criterion variable external to the test itself.

Another use is as a prescriptive aid where different scores imply different treatments. For example, elementary school pupils may be classified according to reading ability on the basis of test scores; they may be assigned to different books or to different kinds of instruction. Job applicants may be classified as marginally employable and assigned to programs of remedial vocational training on the basis of test scores. Disturbed persons with one profile of scores may be assigned to treatments different from those for people with different profiles. Each of these examples implies a hypothesis that people with a specific set of attributes will perform a task or achieve a goal more effectively with one form of treatment than with another. The test user in an applied setting may not have the power, the resources, or the training to carry out the necessary experimental work for testing these hypotheses; he may simply accept them as part of the prevailing scientific or professional body of knowledge and use tests accordingly.

A test score may be used to certify that an individual has met some designated standard, that a person is qualified to perform certain skilled tasks, that a child is qualified for a remedial program, or that a defendant can stand trial.

The basic use of tests is descriptive or evaluative. A test score provides a description of the individual who obtained it and can help the test user to understand, analyze, or help that individual. Test scores may be used by a counselor to help a student make a vocational choice or to help a couple in marriage counseling communicate more clearly with each other. They may help a teacher work more effectively with a pupil. These are clinical, diagnostic, and individualistic uses of tests in a continuing relationship between a test user and an individual. Because the relationship is a continuing one, tentative decisions or judgments can be modified as new information is accumulated.

Test scores may constitute the dependent variable or criterion measure in an institutional research study. A program may be continued or terminated on the basis of test results; an institution may receive more or less funding because of test results; test results may be considered in organizational analysis or in making program changes. These *Standards* do not deal fully with these problems; their emphasis is more on the interpretation of scores of individuals. A companion volume is planned dealing with standards for test use in program evaluation, policy-related research, and

HE USE OF TESTS

desi ated standard, that a person is qua ed to perform certain skilled tasks, that a child is qualified for a emedial program, or that a defendant can and trial.

The basic use of tests is descriptive or evaluative. A test score provides a desc tion of the individual who obta ed it and can help the test user to understand, analyze, or help that individual. Test scores may be used by a cou lor to help a student make a voca nal choice or to help a couple in marriage counseling communicate nor learly with each other. They may help teacher work more effectively with a pupil. These are clinical, diagnostic, and individualistic uses of ests in a continuing relationship etw n a test user and an individual. Because the relationship is a continuing ne, tentative decisions or judgments an modified as new information is ccu ulated.

Test scores may constitute the epe ent variable or criterion measure n a institutional research study. A rogram may be continued or erminated on the basis of test results; n in tution may receive more or less und because of test results; test esults may be considered in rganizational analysis or in making rog n changes. These *Standards* do ot d fully with these problems; their mphasis is more on the interpretation f sc s of individuals. A companion lu is planned dealing with andards for test use in program valuation, policy-related research, and curriculum evaluation; it will also address issues of research design and of data analysis.

The standards in the present volume are to varying degrees directed to all forms of use. As the use of tests moves along a continuum from the description of a single individual, in a situation allowing for corrections of erroneous interpretations, making decisions about large numbers of people, the test user must apply more of the standards and, perhaps, apply them more rigorously. Such decisions may profoundly influence the lives of those tested, such as decisions for employment or for attendance at college, or decisions to assign a person to one treatment or opportunity rather than to another (e.g., tracking in a school system), or decisions to continue or terminate a program or to regulate its funds. The cost of error, in money and in human suffering, may be great. A test user cannot abdicate the responsibilities described in these standards by subscribing to external testing services or test suppliers.

The standards of test use may not have to be so rigidly followed when the purpose of testing is the understanding of an individual. Sometimes such testing is less standardized than is usually recommended. For example, a school counselor may be interested in assessing the maximum performance capability of a single student. To get a full understanding of that student, he must be able to elicit new information, perhaps even through an embellishment of a standardized test, to seek the broadest possible understanding of the level of mastery and of the generalizability of the situations in which mastery can be demonstrated. Interpretation of test scores in such cases is not made in terms of norms but in terms of a counselor's analysis of what mastery of a particular skill entails (even a social skill, not likely to be measured by tests ordinarily used for content-referenced interpretations). In short, exploration of an individual case is different from standardized testing. The *user who develops test embellishments must know the difference;* that is, he must have a clear rationale for what he is doing when he departs from standard procedure, and he must be able to apply that rationale consistently and sensibly. Such individualized testing does not require less skill than does testing broadly for institutional decisions; it requires a different kind of skill (cf. I1).

The standards necessary for using tests for making decisions are not different from the standards necessary when tests are used simply for understanding, but the emphasis within a standard may be different. A test user should be familiar with the standards governing test use in general, and he should pay particular attention to those standards most nearly fitting his own specific type of application.

In doing so, he should realize that the standards are intended to apply, in principle, to *all forms* of assessment. In choosing from alternative methods of assessment, the test user should consider the differences in the ease of applying these standards.

G. Qualifications and Concerns of Users

Assessing others is an occupational activity for teachers, parents,

clergymen, shopkeepers, correction officers, etc. Some people assess with remarkable skill; others are inept and have little or no training to help them. Users of educational and psychological tests in schools, places of employment, clinics, laboratories, prisons, and other places where educators and psychologists work should have had at least some formal training.

A test user, for the purposes of these standards, is one who chooses tests, interprets scores, or makes decisions based on test scores. He is not necessarily the person who administers the test following standard instructions or who does routine scoring. Within this definition, the basic user qualifications (an elementary knowledge of the literature relating to a particular test or test use) apply particularly when tests are used for decisions, and such uses require additional technical qualifications as well. A recurring phrase in discussions about testing is "the legitimate uses of a test." One cannot competently judge whether his intended use is among those that are "legitimate" (however defined) without the technical skill and knowledge necessary to evaluate the validity of various types of inferences.

G1. A test user should have a general knowledge of measurement principles and of the limitations of test interpretations. Essential

[Comment: The required level of knowledge will vary with the complexity of the evaluations to be made and the responsibility of the user. At a minimum, the user must be knowledgeable about testing principles, understand the concept of measurement error, and be able to interpret an obtained test score. He should realize that there are alternative explanations for a given score and should have a pool of knowledge from which to evaluate some of the alternatives.]

G1.1. A test user should know his own qualifications and how well they match the qualifications required for the uses of specific tests. Essential

G2. A test user should know and understand the literature relevant to the tests he uses and the testing problems with which he deals. Very Desirable

[Comment: A broad connotation is intended for this standard. The test user should have some acquaintance with the relevant findings of behavioral sciences, such as those related to the roles of heredity and environment, when using aptitude tests; some understanding of physiology is useful when one is using tests of motor skills. A very narrow interpretation of "the literature relevant to the test" is inadequate.

Unfortunately, it seems that ignorance of the literature requires that old information be rediscovered. For over 40 years, for example, it has been known that children with limited or restricted cultural exposure, such as children on canal boats or in isolated mountain communities, make low scores on intelligence tests standardized on more advantaged populations. The point has been made repeatedly in research reports and textbooks. Nevertheless, many black and Spanish-speaking children with limited cultural exposure who receive low scores on intelligence tests standardized on more advantaged

interpret an obtained test score. He should realize that there are alternative explanations for a given score and should have a pool of knowledge from which to evaluate some of the alternatives.]

G1.1. A test user should know his own qualifications and how well they match the qualifications required for the uses of specific tests. Essential

G2. A test user should know and understand the literature relevant to the tests he uses and the testing problems with which he deals. Very Desirable

[Comment: A broad connotation is intended for this standard. The test user should have some acquaintance with the relevant findings of behavioral sciences, such as those related to the roles of heredity and environment, when using aptitude tests; some understanding of physiology is useful when one is using tests of motor skills. A very narrow interpretation of "the literature relevant to the test" is inadequate.

Unfortunately, it seems that ignorance of the literature requires that old information be rediscovered. For over 40 years, for example, it has been known that children with limited or restricted cultural exposure, such as children on canal boats or in isolated mountain communities, make low scores on intelligence tests standardized on more advantaged populations. The point has been made repeatedly in research reports and textbooks. Nevertheless, many black and Spanish-speaking children with limited cultural exposure who receive low scores on intelligence tests standardized on more advantaged groups are improperly classified as mentally retarded.]

G3. One who has the responsibility for decisions about individuals or policies that are based on test results should have an understanding of psychological or educational measurement and of validation and other test research. Essential

[Comment: A test user should have acquired the technical understanding appropriate to his responsibilities. Test users within organizations or regulatory agencies should have enough technical knowledge to be able to evaluate competently the tests and testing procedures relevant to the decisions they must make. If their technical training is limited, they should seek refresher training or work under the guidance of another test user whose training is adequate.]

G3.1. The principal test users within an organization should make every effort to be sure that all those in the organization who are charged with responsibilities related to test use and interpretation (e.g., test administrators) have received training appropriate to those responsibilities. Essential

[Comment: Serious misuse and distortion in interpretation may occur when people are not properly trained to carry out their responsibilities. The level of training needed varies with the complexity of a testing program, the level of the individual's responsibility for it, and the nature and intensity of possible adverse consequences. Test users should provide at least a basic orientation for administrators or executives who decide whether to test or not to test, to approve or to disapprove specific assessment procedures, to appropriate funds for necessary research, or to decide how test or research results will be used in the organization. Similar knowledge is needed by compliance officers who may have a detrimental influence on testing programs because of unreasoned and unreasoning demands for interpretation of data, who might disapprove of a testing program without adequate consideration of the alternatives, or who might approve faulty and unfair uses of tests out of ignorance.]

G3.1.1. A test user should have sufficient technical knowledge to be prepared to evaluate claims made in a test manual. Very Desirable

[Comment: A test user must accept some responsibility for the choice when a test is chosen. The user must also be able to exercise some judgment concerning descriptions of intended populations that appear in a manual. If he is using the test to evaluate a remedial program for low-performing pupils in the fourth grade, it is not necessarily appropriate to select a test standardized on "children in Grades 4 through 6."]

G3.2. Anyone administering a test for decision-making purposes should be competent to administer that test or class of tests. If not qualified, he should seek the necessary training regardless of his educational attainments. Essential

[Comment: Some tests are easily administered, and a brief explanation of the instructions and of the necessity for standardization may be sufficient training for administering them. The

use of other tests or assessment procedures may require more specific or unique kinds of training, for example, individually administered intellectual or personality measures or some work samples. It should be recognized that the administration and scoring of a test may not require any specific academic degree; conversely, possession of a degree is not necessarily evidence of qualifications to administer a particular test.]

G4. Test users should seek to avoid bias in test selection, administration, and interpretation; they should try to avoid even the appearance of discriminatory practice. Essential

[Comment: This is a difficult standard to apply. Sources of item or test bias are neither well understood nor easily avoided. The very definition of bias is open to question. The competent test user will accept the obligation to keep abreast of developments in the literature and, at the very least, to demonstrate a sensitivity to the problem and to the feelings of examinees.]

G5. Institutional test users should establish procedures for periodic internal review of test use. Essential

[Comment: The competent use of tests and test scores requires regular review of procedures and of concepts that may change with the advent of new knowledge. A practice that might have been considered acceptable or appropriate at an early period may be found to be either harmful or ineffective in light of subsequent findings in psychometric theory or criticisms of test use. The review should examine the soundness of procedures used in test administration, the modes of assessment, the bases for inferences drawn from test scores, and the relative quality of various validation strategies.]

H. Choice or Development of Test or Method

Standardized tests constitute one class of assessment procedures available to the user. He may also choose various kinds of ratings, personal history information, reference information, or "unobtrusive measures." He may also elect to develop his own tests. His choice depends upon what is available for assessing the characteristics of concern, ethical considerations, and his own knowledge and competency. Among standardized tests there are usually many alternatives, different dimensions to be measured, different methods of measurement, and different forms of tests. Choices should be made as deliberately and carefully as circumstances permit; test users should not use habitually the same test or method of assessment for all purposes; neither should they assess only those characteristics that are easily or conveniently assessed and fail to consider other, possibly more important, characteristics. Standards refer to the process of choice, not to the choices themselves.

H1. The choice or development of tests, test batteries, or other assessment procedures should be based on clearly formulated goals and hypotheses. Essential

[Comment: There is usually an assumption that one's goals are good, and that the method of assessment

administration, the modes of assessment, the bases for inferences drawn from test scores, and the relative quality of various validation strategies.]

H. Choice or Development of Test or Method

Standardized tests constitute one class of assessment procedures available to the user. He may also choose various kinds of ratings, personal history information, reference information, or "unobtrusive measures." He may also elect to develop his own tests. His choice depends upon what is available for assessing the characteristics of concern, ethical considerations, and his own knowledge and competency. Among standardized tests there are usually many alternatives: different dimensions to be measured, different methods of measurement, and different forms of tests. Choices should be made as deliberately and carefully as circumstances permit; test users should not use habitually the same test or method of assessment for all purposes; neither should they assess only those characteristics that are easily or conveniently assessed and fail to consider other, possibly more important, characteristics. Standards refer to the process of choice, not to the choices themselves.

H1. The choice or development of tests, test batteries, or other assessment procedures should be based on clearly formulated goals and hypotheses. Essential

[Comment: There is usually an assumption that one's goals are good, and that the method of assessment

chosen will help one achieve those goals. In choosing or building a test one should be able to articulate such assumptions and values. As a general rule, the assumptions take the form of at least an implicit hypothesis: "If I come to a clearer understanding of this individual, in terms of the characteristic or set of characteristics assessed, I will be able to infer something about his vocational success, or his academic problems, or his prognosis in marriage, or whatever."

The use of a test in a decision context implies a hypothesis of the form that a designated outcome is a function of the test variable. A test user should be able to state clearly the desired outcome, the nature of the variables believed to be related to it, and the probable effectiveness of alternative methods of assessing those variables.

The purpose of administering a test should be explicit. In some school systems, it has been alleged, tests are routinely administered with no purpose other than an apparent hope that they will someday be useful. Such routine testing is unwise.]

H1.1. The test user should formulate goals clearly enough so that he can later evaluate his success in achieving them and can communicate that evaluation to other qualified persons. Very Desirable

[Comment: In a clinical or counseling situation, there is a continuing relationship with a person who has been tested. If a test user has clear purposes, later events can provide evidence of some success in achieving goals or information for changing inferences or procedures. Without a clear idea of what is to be learned about a person, and why, there will be no clear direction in the counseling relationship.

One's purposes in developing a testing program define his criteria, and the nature of the criteria should suggest to the informed user hypotheses, that is, test variables, that might be associated with them. Such hypotheses should be reasonable. There is no clear reason, for example, to use a mechanical-aptitude test to try to predict performance in English classes. Similarly, it is not easy to see what purpose is intended when a vocabulary test is adopted for use in the selection of rolling-mill employees.

Some hypotheses are much more easily justified than others. Few people will quarrel with the suggestion that applicants who type rapidly with few errors will become preferred employees in a stenographic pool. The hypotheses that those who are likely to work with greater persistence at a routine, manipulative task can be identified by scores on a very long but easy perceptual speed and accuracy test may require a more detailed explanation of the logic and background data.]

H1.2. The test user should consider the possibility that different hypotheses may be appropriate for people from different populations. Essential

H2. A test user should consider more than one variable for assessment and the assessment of any given variable by more than one method. Essential

[Comment: For most purposes, the evaluation of a person requires description that is both broad and precise; a single assessment or assessment

procedure rarely provides all relevant facets of a description.

Decisions about individuals should ordinarily be based on assessment of more than one dimension; when feasible, all major dimensions related to the outcome of the decision should be assessed and validated. This is the principle of multivariate prediction; where individual predictors have some validity and relatively low intercorrelations, the composite is usually more valid than prediction based on a single variable. It is not always possible to conduct the empirical validation study (certainly not in working with problems of individuals one at a time), but the principle can be observed.

In any case, care should be taken that assessment procedures focus on important characteristics; decisions are too often based on assessment of only those dimensions that can be conveniently measured with known validity. For example, mental retardation is often defined as both deficiency in tested intelligence and poor adaptive behavior. If both parts of this definition are accepted, then both variables should be considered in deciding whether an individual is to be classified as a mental retardate, even though it is much more difficult to measure adaptive behavior than to find an acceptable scale for testing intelligence.

Test users should also consider more than one method of assessment. Even a test yielding generally valid scores may in an individual case be susceptible to idiosyncratic errors of interpretation, and a pattern of confirming or modifying assessments may be useful. Confidence in inferences drawn from assessments may be increased by varying the sources and increasing the amount of information on which the inferences are made. In addition to tests, one might consider ratings, references, observations of actual performance, etc. Of these, a test is probably most valid. If the others add to the validity of an assessment, they should be systematically considered in statistical prediction; otherwise, they should be ignored. Frequently, however, one will not have enough confidence in test interpretations to justify overlooking other data. In particular, when using a given test with minorities, one may question the validity of test inferences for those populations and want to get as much additional information as possible before making decisions.]

H2.1. In choosing a method of assessment, a test user should consider his own degree of experience with it and also the prior experience of the test taker. Essential

[Comment: Inexperience of the test user can be alleviated by reading, practice, and training. Warm-up tests or other methods of acclimatization are advocated to alleviate the inexperience of test takers. In addition, attention should be given to the degree of interaction between test user and test taker; there may be special sources of anxiety in situations where they are of different cultural or ethnic background.]

H2.2. The choice or development of a test or test procedure, or the addition of a test or test procedure to existing assessments, should involve consideration of the relationship between the cost of the choice and the benefit expected. Very Desirable

varying the sources and increasing the amount of information on which the inferences are made. In addition to tests, one might consider ratings, references, observations of actual performance, etc. Of these, a test is probably most valid. If the others add to the validity of an assessment, they should be systematically considered in statistical prediction; otherwise, they should be ignored. Frequently, however, one will not have enough confidence in test interpretations to justify overlooking other data. In particular, when using a given test with minorities, one may question the validity of test inferences for those populations and want to get as much additional information as possible before making decisions.]

2.1. In choosing a method of assessment, a test user should consider his own degree of experience with it and also the prior experience of the test taker. Essential

[Comment: Inexperience of the test user can be alleviated by reading, practice and training. Warm-up tests or other methods of acclimatization are advocated to alleviate the inexperience of test takers. In addition, attention should be given to the degree of interaction between test user and test taker; there may be special sources of anxiety in situations where they are of different cultural or ethnic background.]

.2. The choice or development of a test or test procedure, or the addition of a test or test procedure to existing assessments, should involve consideration of the relationship between the cost of the choice and the benefit expected. Very Desirable

[Comment: Both costs and benefits may involve broader considerations than the specific problem at hand. Although quantification may be difficult and even unreliable, costs and benefits to the individuals tested and to the broader society should enter into consideration.]

H3. In choosing an existing test, a test user should relate its history of research and development to his intended use of the instrument. Essential

[Comment: A school system was faced with the necessity of reducing its faculty. Reductions in force, according to policy, were to be based on teacher competence. However, decisions were in fact based on scores on a test that had been developed to evaluate the educational backgrounds of new teachers-college graduates. Nothing about the test established its validity as a measure of classroom effectiveness, nor was any local research conducted on this point. Its choice was, therefore, inappropriate.

In a different situation, a decision had been made to use a standard achievement test to evaluate pupil progress. Upon investigation of the test's development, it was found inconsistent with the curriculum objectives of that school. Other tests were examined and an alternative test was chosen that more closely matched the curriculum content. (In some cases, closely matching curriculum content may not be advantageous since it prevents one from knowing the extent to which pupils may be deficient in skills or knowledge not deliberately specified in the local curriculum objectives.)

This standard calls for a general evaluation of the validity of the proposed use of a test. Such an evaluation includes evaluation of the procedures followed in the development of the test and of the quality and relevance of the research that has been done with it.]

H4. In general a test user should try to choose or to develop an assessment technique in which "tester-effect" is minimized, or in which reliability of assessment across testers can be assured. Essential

[Comment: In general, the less the influence of the tester on scores, the fairer the test. The influence of the tester is obviously greater in an unstructured interview than in a structured one, and there may be more tester effect in a structured interview than in a structured personal history form. Tester effect is most likely to be minimized by standardized testing. In using tests, some organizations have turned to tape-recorded instructions in an effort to minimize further possible tester effect.]

H5. Test scores used for selection or other administrative decisions about an individual may not be useful for individual or program evaluation and vice versa. Desirable

[Comment: The purposes of institutional testing and of evaluative testing are not always compatible. Whereas the typical evaluative use is intended to help the individual (or program), institutional decisions frequently have the apparent effect of hurting, even if the decision may have unseen long-term benefits to the individual (such as avoidance of an un-

necessary failure experience). These seemingly contradictory functions may prevent effective interpretation in either instance.]

I. Administration and Scoring

A test user may delegate to someone else the actual task of administering or scoring tests, but he retains the responsibility for these activities. In particular, he has the responsibility for ascertaining the qualifications of such agents. Standards for administration apply not only to the act of testing but also to more general matters of test administration. The basic principle is standardization; when decisions are based on test scores, the decision for each individual should be based on data obtained under circumstances that are essentially alike for all.

I1. A test user is expected to follow carefully the standardized procedures described in the manual for administering a test. Essential

[Comment: It may in rare cases be necessary to modify procedures. When there is any deviation from standard practice, it should be duly noted, and interpretations of scores should not be made in terms of normative data provided in the manual. Modifications may be standardized for specific purposes. For example, modifications were necessary for testing a group of deaf mutes, but the modified procedures were applied in the same way to all children in the group.

Standardization of procedure is particularly important when decisions are based on test scores. Nevertheless, known examples of failure to follow standardized procedures are numerous. In one organization, when test supplies were depleted, mimeographed versions of uneven quality were prepared. (In addition to being a violation of copyright laws, the result was a markedly changed set of stimulus materials). One test administrator tried to relieve the monotony of the repeated verbatim instructions by giving a shorter and varied version.]

I1.1. A test user must fully understand the administrative procedures to be followed. Essential

[Comment: The test user should be fully trained to do whatever is required for competent administration of the test. The administration of many tests requires nothing more than the ability to read and understand routine instructions, that of others requires extensive supervised practice. Whatever the requirement, the test user lacking such training should seek it.]

I1.2. A test user should maintain consistent conditions for testing. Very Desirable

[Comment: Situational variables should be reasonably controlled. For example, there should be no great variation in temperature or humidity; noises and other distractions should be as nearly eliminated as possible. Interadministrator reliability should be high. In general, testing conditions should minimize variations in the testing procedure.]

I1.3. A test user should make periodic checks on material, equipment, and procedures to maintain standardization. Essential

standardized procedures are numerous. In one organization, when test supplies were depleted, mimeographed versions of uneven quality were prepared. (In addition to being a violation of copyright laws, the result was a markedly changed set of stimulus materials). One test administrator tried to relieve the monotony of the repeated verbatim instructions by giving a shorter and varied version.]

I1.1. A test user must fully understand the administrative procedures to be followed. Essential

[Comment: The test user should be fully trained to do whatever is required for competent administration of the test. The administration of many tests requires nothing more than the ability to read and understand routine instructions, that of others requires extensive supervised practice. Whatever the requirement, the test user lacking such training should seek it.]

I1.2. A test user should maintain consistent conditions for testing. Very desirable

[Comment: Situational variables should be reasonably controlled. For example, there should be no great variation in temperature or humidity; noises and other distractions should be as nearly eliminated as possible. Interadministrator reliability should be high. In general, testing conditions should minimize variations in the testing procedure.]

I1.3. A test user should make periodic checks on material, equipment, and procedures to maintain standardization. Essential

[Comment: This standard refers, for example, to the constancy of graphics and of printing, or the accuracy of stop watches. It also seeks the elimination of bad habits that may creep into administrative procedures. It applies particularly to any testing procedures that make use of physical equipment which is subject to wear. Such equipment should be regularly examined to assure that its characteristics remain within acceptable tolerances; for example, a pegboard should be replaced if holes become enlarged or beveled through use.]

I2. The test administrator is responsible for establishing conditions, consistent with the principle of standardization, that enable each examinee to do his best. Essential

[Comment: In a negative sense, the goal of this standard is that conditions inhibiting maximum performance should be avoided. The principle can be followed in part simply by being sure that all materials—such as answer sheets, pencils, and erasers—are on hand and that precautions have been taken to avoid distractions. In a more positive sense, the administrator should be sure that the examinee understands the tasks involved in taking the test: what kinds of responses are to be made and on what answer sheets, the implications for test-taking strategy of erasures or multiple marking or guessing, and how to know whether the test has been completed.

The tester should try to create a nonhostile environment; standardized procedures are impersonal, but the test administrator must avoid being either patronizing or unresponsive to the examinees, especially when the tester and the examinee differ in race, sex, or status. A testing situation contains elements that are nonrecurring and unique to the persons tested. Although these may have negligible effects on test reliability, they may include events perceived as denigrating or questioning of the worth of the individual. A complete catalog of such events is not possible or easily described. In general, however, the social amenities of respect, politeness, and due regard for extenuating circumstances are relevant guides for insuring the dignity of persons. While it may not be demonstrated that abuse of these principles leads to poor test performance, such abuse is not likely to enhance performance.

It is often difficult to maximize the motivation of the examinees. The attempt is important; a major source of error may arise when examinees do not like or trust the test, tester, or test situation, and therefore make no special effort to do well in it.]

I2.1. Procedures manuals should be prepared for use in organizations when there is repeated testing. Very Desirable

[Comment: Just as a manual is needed for a test, a manual is needed for a testing program. Changes in personnel or lapses in memory make a record of procedures developed and followed necessary for standardization. Such a manual might indicate appropriate circumstances for testing or for referral for testing, standard sequences of tests, or guides to interpretations of test batteries in addition to instructions for administering and scoring tests taken or adapted from individual test manuals.]

I3. A test user is responsible for accuracy in scoring, checking, coding, or recording test results. Essential

[Comment: Any agent of the user shares this responsibility. The clerk who scores a test must understand and accept the necessity for accuracy. The test user, who may or may not do the actual scoring, nevertheless has the responsibility to be sure that procedures are established and followed for verifying accuracy. It is unfair to individuals or organizations when decisions are based on avoidable error.]

I3.1. When test scoring equipment is used, the test user should insist on evidence of its accuracy; when feasible, he should make spot checks against hand scoring or develop some other system of quality control. Essential

[Comment: The frequency of such checks will depend on what is known of the procedures on checking within the scoring service. Commercial scoring services may be queried about their procedures if they have not already announced them; if the procedures seem well designed, such spot checks may be needed only infrequently. Some computer services, on the other hand, may be less meticulous, and some hand scoring may be required in each batch of tests scored by machine. One test user in a certification program, where machine analysis of answers not only yields individual scores but also supplies data for analysis prior to revisions, has adopted the policy of hand scoring as well as machine scoring each test. This assures the accuracy of every score used in individual decisions, and it also assures the accuracy of the computer data used in the continuing research program.]

I3.2. When test scoring requires judgment, the test user should determine interscorer or intrascorer reliability. Very Desirable

[Comment: When the test user does his own scoring, he should make periodic comparisons of scores he has determined against scores on the same sets of responses determined by other scorers or by himself at other times.]

I4. If specific cutting scores are to be used as a basis for decisions, a test user should have a rationale, justification, or explanation of the cutting scores adopted. Essential

[Comment: When a cutting score is adopted, the effect is to reduce scoring to a scale of only two points: pass and fail. The validity of the test scored in this way is different at different cutting scores and, in general, is different from the validity found with continuous scores.

The test user should have some justifiable reason for the adoption of a given cutting score. Many kinds of arguments might be used. In a content-referenced interpretation of a mastery test, such a score might be determined as the obtained score at which one can reject, at a preselected level of probability, the hypothesis that a predesignated confidence interval for that score includes the perfect score on the test. If interpretations are referenced against an external criterion, the cutting score might be one where there is a designated probability of achieving a specified level of success (e.g., "We do not admit students who have less than a

DS

a used in the continuing research gram.]

3.2. When test scoring requires gment, the test user should determe interscorer or intrascorer abi. Very Desirable

Comment: When the test user does own scoring, he should make iodi comparisons of scores he has erm ed against scores on the same s of responses determined by other rers or by himself at other times.]

4. **specific cutting scores are to used as a basis for decisions, a test r should have a rationale, tific on, or explanation of the cut- sc s adopted.** Essential

Comment: When a cutting score is pted the effect is to reduce scoring a sc of only two points: pass and . The validity of the test scored in s way is different at different cutting res d, in general, is different from va ity found with continuous res.

The test user should have some tifi e reason for the adoption of a en tting score. Many kinds of uments might be used. In a contentere d interpretation of a mastery t, s a score might be determined the obtained score at which one can ect, at a preselected level of probility the hypothesis that a presign d confidence interval for that re includes the perfect score on test. If interpretations are referenced ains an external criterion, the ting core might be one where there a designated probability of achieving pec d level of success (e.g., "We do ad t students who have less than a 30 per-cent chance of graduating"). Decision theory principles can be used to find a cutting score that will maximize the discrimination between high- and low-criterion groups. One might base the cutting score simply on a distribution of scores in a "predicted-yield" situation; for example, the proportion of job applicants who accept offered employment, the number of new employees who will be needed, predictions of growth or reduction in force, and related information can be used to determined the "predicted yields" of new employees at different cutting scores. The determination of a cutting score on this basis may result in using the test in a range where it is less than maximally effective; its validity should be determined in light of its actual use.

This standard does not attempt to recommend a specific procedure for developing cutting scores where they are to be used. The intent is to recommend that test users avoid the practice of designating purely arbitrary cutting scores they can neither explain nor defend. Cutting scores adopted with reference to those used in another organization or for another purpose, or by a casual glance at normative tables, are usually unsatisfactory.]

I4.1. A validity coefficient for each criterion for which a predictor test is recommended should be provided at each one of several points on the score continuum that may be used as cutting scores. Very Desirable

I4.1.1. If examinees are to be selected on the basis of a set of scores that displays different regression lines for use in predicting the same criterion in different subgroups of an applicant population, cutting scores should be established with great caution to avoid unfairness to members of one or more of the subgroups. Essential

[Comment: There are many defensible definitions of "unfairness" in the literature, and techniques have been developed for setting cutting scores to minimize "unfairness" as defined in these ways. Test users should keep abreast of the rapidly developing literature on this topic.]

I5. The test user shares with the test developer or distributor a responsibility for maintaining test security. Essential

[Comment: Test security is a problem whenever a lapse in security can result in changing an individual's score without making a change in his true score. For some kinds of tests a lapse of security would not be serious. If one is to be tested for achieved skill, for example, knowing and practicing the test samples might be highly recommended. In many cases, however, prior knowledge of test items or scoring procedures could destroy validity. The problem is not simply one of cheating. Security may be compromised where examinees have had much prior experience with a popular test, have been taught specific test items, or have heard a lot about the test.

I5.1. Where a probable breach of security may invalidate test inferences, the test user should employ other methods of assessment; that is, he should seek a basis for more valid inferences. Very Desirable

I5.2. All reasonable precautions

should be taken to safeguard test material. Essential

[Comment: The use of locked files is a minimal requirement in maintaining test security. It is important to know the recipient whenever tests are out of the filing cabinets. In a test-taking situation, examinees should be proctored. When a test is mailed out to other locations, the recipient should be known and trusted. The ubiquitous copying machine has intensified the problem of safeguarding test materials.]

I5.3. The test user should avoid basing decisions on scores obtained from insecure tests. Very Desirable

[Comment: A test may be designated insecure because it is known that unauthorized copies have gone astray. Another test might be considered insecure because it is so widely used that a test taker may have had ample opportunity to practice it in other test-taking situations previously and be able to recognize items. Some employment tests, for example, are so widely used that a job applicant may have taken them several times while applying to various employers.]

J. Interpretation of Scores

Standards in this section refer to the interpretation of a test score by the test user and to reports of interpretations. Reports may be made to the person tested, to his agent, or to other affected people: Teachers, parents, supervisors, and various administrators and executives.

J1. A test score should be interpreted as an estimate of performance under a given set of circumstances. It should not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances. Essential

J1.1. A test user should consider the total context of testing in interpreting an obtained score before making any decisions (including the decision to accept the score). Essential

[Comment: The standard is that one must avoid the abdication of responsibility by relying exclusively on an obtained score. Users should, in particular, look for contaminating or irrelevant variables that may have influenced obtained scores; for example, in testing to classify school children, scores may be influenced by behavior problems, visual or hearing defects, language problems, and racial or cultural factors, as well as by ability.]

J2. Test scores should ordinarily be reported only to people who are qualified to interpret them. If scores are reported, they should be accompanied by explanations sufficient for the recipient to interpret them correctly.

[Comment: There are difficult problems associated with the question of who should have access to test scores within an organization. Certainly, curious peers should not have access to them. An individual who must make the ultimate decision to admit or to reject or to hire or to reject, or to certify or not to certify, must have the interpretation. One useful (and unanswered) question is whether such a person who lacks the training necessary for the interpretation of scores should be given that training or should be given only the interpretations of scores.]

ould not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances. sential

J1.1. A test user should consider the tal context of testing in interpreting obtained score before making any cisions (including the decision to accept the score). Essential

Comment: The standard is that one st avoid the abdication of responsibility by relying exclusively on an obtained score. Users should, in particular, look for contaminating or elevant variables that may have influenced obtained scores; for example, testing to classify school children, res may be influenced by behavior problems, visual or hearing defects, language problems, and racial or tural factors, as well as by ability.]

2. Test scores should ordinarily be reported only to people who are qualified to interpret them. If scores are orted, they should be accompanied by explanations sufficient for the recipient to interpret them correctly.

Comment: There are difficult blems associated with the question of who should have access to test scores hin an organization. Certainly, ious peers should not have access to em. An individual who must make the ultimate decision to admit or to ect or to hire or to reject, or to certify not to certify, must have the interpretation. One useful (and unanswered) question is whether such a per who lacks the training necessary for interpretation of scores should be given that training or should be given y the interpretations of scores.]

J2.1. An individual tested (or his agent or guardian) has the right to know his score and the interpretations made. In some instances, even scores on individual items should be made known. Desirable

[Comment: Strictly speaking, this is an ethical standard rather than a standard of competent test use; it is stated here because it conflicts with technical considerations of test security. If the standard is followed, test interpretations and their foundations will be made available to those with a "need to know"; certainly, the individual whose future is affected by the decision is among those with a "need to know." The test user should take any precautions he can, when the demand for information is severe, to protect test security, but he should not do so at the expense of an individual's right to understand the bases for decisions that affect him adversely. Such understanding may be better promoted, with less threat to test security, by using qualified persons sympathetic to the individual's interests. For example, when there is a civil-rights issue, it would be most useful to have items examined by a qualified testing specialist who is known to be both concerned and knowledgeable.]

J2.2. A system of reporting test results should provide interpretations. Essential

[Comment: Although the form of a report will differ for different audiences (examinees, teachers, parents, supervisors), it should communicate the interpretation in a form that will be clear and easily understood.]

J2.2.1. Scores should ordinarily be interpreted in light of their confidence intervals rather than as specific values alone. Very Desirable

J2.3. In general, test users should avoid the use of descriptive labels (e.g., retarded) applied to individuals when interpreting test scores. Desirable

[Comment: The standard applies to the use of summary diagnoses in general. For nearly all purposes, it is better to describe behavior and to differentiate such description from inference. Summary labels tend to be inferences treated as if they were descriptions.

The use of a summary label generally connotes value judgments; unfortunately, most are words used in everyday language and therefore subject to inaccurate interpretation. A test maker may know precisely what he means when he uses the term "retarded," but he has no influence over the interpretation of the same word by a judge, teacher, parent, or child.]

J3. The test user should recognize that estimates of reliability do not indicate criterion-related validity. Essential

[Comment: Reliability is a necessary but not a sufficient condition of validity. Reliability coefficients are pertinent to validity in the negative sense that unreliable scores cannot be valid; but reliable scores are by no means *ipso facto* valid.]

J4. A test user should examine carefully the rationale and validity of computer-based interpretations of test scores. Essential

[Comment: The user of a special service has the obligation to be thoroughly familiar with the principles on which such interpretations are derived, and he should have the ability to evaluate a computer-based interpretation of test performance in light of other evidence he may have.

J5. In norm-referenced interpretations, a test user should interpret an obtained score with reference to sets of norms appropriate for the individual tested and for the intended use. Essential

[Comment: The reverse is also a standard of competent test use: The test user ordinarily should not interpret an obtained score with reference to a set of norms that is inappropriate for the individual tested or for the purposes of the testing. This is a relatively simple standard to state, but it often is difficult to apply. Contemporary social problems suggest that men and women or members of different ethnic groups should for some purposes be evaluated in terms of several norms groups. For other purposes, such as vocational counseling, students should know how they stand relative to those in or entering a relevant occupation, regardless of their ethnic background. Of course, women or members of minority groups should not be counseled to avoid nontraditional occupations (e.g., women in engineering) merely for lack of appropriate norms.

It is by no means certain that sex or race is the crucial variable in interpreting a given score. It may well be that more important variables for differential norms would be breadth of cultural exposure (or degree of cultural isolation), skill and experience in the use of standard English, interests, or similar variables which may seem to be related to sex or racial differences in test performance.]

J5.1. It is usually better to interpret scores with reference to a specified norms group in terms of percentile ranks or standard scores than to use terms like IQ or grade equivalents that may falsely imply a fully representative or national norms group. Essential

J5.2. Test users should avoid the use of terms such as IQ, IQ equivalent, or grade equivalent where other terms provide more meaningful interpretations of a score. Essential

[Comment: Such scores are objectionable for several reasons. Most important, they generally involve spurious projections of growth. They involve an interpretation which is at best awkward. (To illustrate: It is much simpler to ask, in interpreting a score, "Where does this person stand in relation to specific norm groups?" than to ask, "What group is this person's performance like the average of?" The semantic awkwardness of the latter question illustrates its psychometric awkwardness as well.) They are labels to which the general public attaches many different inappropriate meanings.

Some of these scores, such as mental age or grade equivalent scores, involve severe technical problems. For example, serious misinterpretations occur when grade levels are extrapolated beyond the range for which the test is designed. Moreover, it should be recognized that the standard error of measurement for some widely used

isolation), skill and experience in the use of standard English, interests, or similar variables which may seem to be related to sex or racial differences in test performance.]

J5.1. It is usually better to interpret scores with reference to a specified norms group in terms of percentile ranks or standard scores than to use terms like IQ or grade equivalents that may falsely imply a fully representative or national norms group. Essential

J5.2. Test users should avoid the use of terms such as IQ, IQ equivalent, or grade equivalent where other terms provide more meaningful interpretations of a score. Essential

[Comment: Such scores are objectionable for several reasons. Most important, they generally involve spurious projections of growth. They imply an interpretation which is at best awkward. (To illustrate: It is much simpler to ask, in interpreting a score, "Where does this person stand in relation to specific norm groups?" than to ask, "What group is this person's performance like the average of?" The semantic awkwardness of the latter question illustrates its psychometric awkwardness as well.) They are labels to which the general public attaches many different inappropriate meanings.

Some of these scores, such as mental age or grade equivalent scores, involve severe technical problems. For example, serious misinterpretations occur when grade levels are extrapolated beyond the range for which the test is designed. Moreover, it should be recognized that the standard error of measurement for some widely used standardized achievement tests may be equal to one grade level.

If a test user, either because of his own limitations or because of rigid institutional policies, feels that he must use such terms, he should be sure that interpretations are also given in standard scores or percentile ranks with reference to the specific norms group used in deriving them. The specific test, test form, time of testing, and nature of the test situation should be included in the statement.]

J5.3. A test user should examine differences between characteristics of a person tested and those of the population on whom the test was developed or norms developed. His responsibility includes deciding whether the differences are so great that the test should not be used for that person. Essential

J5.3.1. If no standardized approach to the desired measurement or assessment is available that is appropriate for a given individual (e.g., a child of Spanish-speaking migrant workers), the test user should employ a broad-based approach to assessment using as many methods as are available to him. Very Desirable

[Comment: The standard is to do the best one can. This may perhaps include the use of a test, even though no appropriate normative data are available, simply as a means of finding out how the individual approaches the task of the test. It might include references, extensive interviews, or perhaps some *ad hoc* situational tasks. Efforts to help solve educational or psychological problems should not be abandoned simply because of the absence of an appropriate standardized instrument.]

J5.4. Local normative data or expectancy tables should ordinarily be developed, if possible, when administrative decisions are based on test scores. Very Desirable

[Comment: Expectancy tables may be more useful than norms. When decisions are based on test scores (with the possible exception of content-referenced interpretations), the test user has ordinarily hypothesized that some outcome on an external criterion is related to performance on the test. Decision makers will have a more useful interpretation of a test score if it is expressed in terms of an expected level of performance on the criterion than if it is expressed in terms of relative standing.]

J5.5. Ordinarily, normative interpretations of ability-test scores should not be made for scores in the chance range.

[Comment: On one reading test for elementary school students, a child who cannot read, and therefore gives truly random responses, would be most likely to obtain a grade-equivalent score, according to the norms, of 2.2; that is, second month of second grade. Quite apart from the usual difficulties with grade-equivalent scores, the example demonstrates the impropriety of trying to make a normative interpretation of a test score obtained in a chance range. One test manual for a widely used test of general mental ability has provided a useful guide to the interpretation of "range-of-chance" scores.]

J6. Any content-referenced inter-

pretation should clearly indicate the domain to which one can generalize. Essential

J7. The test user should consider alternative interpretations of a given score. Essential

[Comment: In a sense, a test-score interpretation implies the hypothesis that the score obtained is a function of the trait level "really" possessed. Alternative hypotheses can be suggested. The obtained score might be a function of anxiety, prior knowledge of the test, inadequate understanding of the instructions, a general sort of test wiseness, deliberate faking, or any of several other possibilities. The test user needs to consider more than the obvious interpretation and to have the skill and sensitivity necessary to develop alternative explanations and to evaluate them.]

J7.1. Where cutting scores are established as guides for decision, the test user should retain some degree of discretion over their use. Desirable

[Comment: The point bears repeating that a test user cannot abdicate the responsibility for the decision to use the test. In most circumstances, there *are* alternatives. Despite the fact that a given test may have a high predictive validity for a specific function, it may represent a trait which is not the only path to success in the predicted venture; and its validity for a given individual, tested at a particular time and under particular circumstances, may be in doubt.

This standard may *not* be taken as a license to discriminate; it is to be used sparingly in recognition that excessive subjectivity can reduce rather than enhance validity. The intent is to avoid a mechanical rigidity in using test scores of imperfect validity. See also H2.]

J7.2. A person tested should have more than one kind of opportunity to qualify for a favorable decision. Desirable

[Comment: In some situations, a candidate might be given the option to qualify on the basis of characteristics other than those measured by the test. If a person with a score so low that his best prognosis is academic failure, nevertheless succeeds in college, he may have demonstrated qualities necessary for success other than those measured by the test, and the fact might well be considered.

Again, the standard must be judiciously applied. In general, the most valid methods available should guide decisions; the subjective use of information not validated can reduce validity. When compelling information exists, however, it should not be ignored in individual cases. It should be noted that it would be unethical as well as invalid to invoke this principle in the application of particular biases of the test user.]

J7.3. A procedure for reporting test results should include checks on accuracy and make provision for retesting. Desirable

[Comment: Errors in procedures and in test scoring occur; procedures should be available for checking. Retesting is one form of checking results. There should be some limits to a retesting provision. The number of allowable

subjectivity can reduce rather than enhance validity. The intent is to avoid a mechanical rigidity in using test scores of imperfect validity. See also H2.]

J7.2. A person tested should have more than one kind of opportunity to qualify for a favorable decision. Desirable

[Comment: In some situations, a candidate might be given the option to qualify on the basis of characteristics other than those measured by the test. If a person with a score so low that his best prognosis is academic failure, nevertheless succeeds in college, he may have demonstrated qualities necessary for success other than those measured by the test, and the fact might well be considered.

Again, the standard must be judiciously applied. In general, the most valid methods available should guide decisions; the subjective use of information not validated can reduce validity. When compelling information exists, however, it should not be ignored in individual cases. It should be noted that it would be unethical as well as invalid to invoke this principle in the application of particular biases of the test user.]

J7.3. A procedure for reporting test results should include checks on accuracy and make provision for retesting. Desirable

[Comment: Errors in procedures and in test scoring occur; procedures should be available for checking. Retesting is one form of checking results. There should be some limits to a retesting provision. The number of allowable retests may be limited by the number of parallel forms available. Certain types of assessments (e.g., personal–history data) are inappropriate for "retesting." Moreover, a true score is likely to be closer to the mean than is its corresponding obtained score. This fact has important implications for extremely low scores; they will tend to be increased in a retest.

In general, however, opportunities for retesting should be permitted without major obstacles. The principle is that no one should be a victim, without recourse, of an adverse decision on the basis of faulty and correctable psychological assessment. Nor should such decisions be permanent; over a period of time, individuals should have a chance for reevaluation on the basis of new learning or new experience.]

J8. The test user should be able to interpret test performance relative to other measures. Very Desirable

[Comment: For many uses, one should be able to interpret test scores in terms of external criteria. The necessary data may be in a test manual; manuals for some academic aptitude tests provide expectancy charts useful for such interpretations. When an assessment of performance on an external criterion is also available, and when there is a wide discrepancy between actual and predicted criterion performance, the test user should investigate possible reasons for the discrepancy. Furthermore, there should be no *a priori* assumption that either the test or the criterion is the instrument in error.]

J8.1. A test user should be able to use and interpret data regarding the statistical significance of differences between scores. Very Desirable

[Comment: A test user may observe differences in scores made by two individuals on the same test. There may be differences in the scores made by an individual on a pretest and a posttest after some intervening treatment or training. He may be interested in comparing the individual's performance on one test with the same person's performance on still another test. In such cases, the test user should know how much confidence to place in an observed score difference.]

J9. A test user should develop procedures for systematically eliminating from data files test-score information that has, because of the lapse of time, become obsolete. Essential

[Comment: Data should not even be available for consideration in decision making after an invalidating period of time. Scores on early achievement tests in areas where later learning or forgetting is to be expected (e.g., an old typing-test score) are no longer likely to be valid.

Not all data are equally susceptible to obsolescence. Information of a highly subjective nature might be judged to become obsolete in a shorter time than more objective items of information. Information about young children might be judged to become obsolete in a shorter period of time than comparable kinds of information about adults.

In the case of data that have potential value for research or for survey purposes, the purging may consist of destroying the link between a person's name and the information relevant to the test rather than destroying the information itself.]

INDEX

Achievement tests
  classification of items, E12.2
  content-by-process outline, E12.2.1
  validity, intro. to, E12.2, E12.2.1
Administration of tests, C1ff, I1ff
Answer sheets
  directions to subjects, C.2
  interchangeability of different types, C3.3

Bias in sampling, E9

Central tendency
  norms, D3
  validity of sample, E6.2
Consistency
  internal, F5ff
  overtime, *see* Stability
  see also Reliability
Construct validity, *see* Validity, construct
Contamination of criterion, E4.1, E4.1.1, E7.3
Content validity, *see* Validity, content
Content reference, J6
Corrected coefficients
  of validity, E8.2.1
  of reliability, F3
Correlation
  item score, E1.4
  with other tests measuring the same attribute, E13.1.2
  see also Validity, Regression, Reliability
Criterion data
Cross-validation, E10ff
Data, continue collection, E11
Date
  establishment of content validity, E7.4.1
  testing and collection of criterion data, E7.4, E11
Directions for administration, C1.1, Gff
Directional diagnosis, E8.3.1

Error in test interpretation, B1.4
  see also Interpretation of scores, Measurement error
Experimental or research test, B2.1

Factorial studies, E13.2.2

Group vs. individual interpretation of scores, D7
Guessing, correction for, C3.5
Internal consistency, F5ff
  see also Reliability
Interpretation of scores, B
  case studies used in, B1.3
  checking validity of, E
  content referenced, B3.1
  computer-based, B6, J4
  effect of test names, B1.1ff
  error, B1.3, B1.4
  misuse, prevention of, B1ff
  stability, F6
  subscores, B5.5
Interpretation, aids to, Bff, E1ff
Interpretive scores, D5.2.3
Instruction to subjects, I1
Items
  description of items used, E9
  interpretation, E9, E1.4
  relationship to other items, E8.32, E8.4, E8.5
  writers, A2.4
Item-test correlations, E12.1

Local norms, *see* Norms
Long-term prediction, E7.4.3, E7.4.4

Manual, A1ff
  absence of information, A1.1
  administration of tests, C1ff
  information too extensive for, A1.2
  contents of, A2ff, B1ff
  norms, D1
  specialized tests, E13.2.1
  purpose of, Aff
  revision, A3ff
  short form, A3ff
Measurement error
  correction for, E8.2.1, E13.2.6
  for different groups, F2.3.1, F6.4
  for different score levels, D2.3.2, F6.3
  general principles, F
  inconsistencies, F1.1
  for short form of test, A3.2.1, A3.3.2
  in test interpretation, B1.3, B1.4, Jff
  procedures and samples, F2ff

Names of tests, B1.1ff
Nature of tests, B3ff
Norms, D
  description of norms group, D2ff, D5.4
  local norms, D1.1, D1.2, D4
  method of deriving, D2.1.1, D2.1.2
  revision, A3.2ff
  user interpretation, J5ff

Obsolescence
  scores, J9

checking validity of, E
content referenced, B3.1
computer-based, B6, J4
effect of test names, B1.1ff
error, B1.3, B1.4
misuse, prevention of, B1ff
stability, F6
subscores, B5.5
Interpretation, aids to, Bff, E1ff
Interpretive scores, D5.2.3
Instruction to subjects, I1
Items
description of items used, E9
interpretation, E9, E1.4
relationship to other items, E8.32, E8.4, E8.5
writers, A2.4
Item-test correlations, E12.1

Local norms, *see* Norms
Long-term prediction, E7.4.3, E7.4.4

Manual, A1ff
absence of information, A1.1
administration of tests, C1ff
information too extensive for, A1.2
contents of, A2ff, B1ff
norms, D1
specialized tests, E13.2.1
purpose of, Aff
revision, A3ff
short form, A3ff
Measurement error
correction for, E8.2.1, E13.2.6
for different groups, F2.3.1, F6.4
for different score levels, D2.3.2, F6.3
general principles, F
inconsistencies, F1.1
for short form of test, A3.2.1, A3.3.2
in test interpretation, B1.3, B1.4, Jff
procedures and samples, F2ff

Names of tests, B1.1ff
Nature of tests, B3ff
Norms, D
description of norms group, D2ff, D5.4
local norms, D1.1, D1.2, D4
method of deriving, D2.1.1, D2.1.2
revision, A3.2ff
user interpretation, J5ff

Obsolescence
scores, J9

tests and interpretation, Iff
Organizational use of material, A1.2.3

Percentile scores, D3, D5.3
Personality inventories
names, B1.1
Procedures, standardization of, Cff
Promotional material for tests, A1.2.1, A1.2.2
Psychiatric diagnosis
agreement with a test, E4.4.2, E4.4.3
Purposes of tests, B2ff

Qualifications
test administrators, B4ff, G3.1
see also Users
Quantitative information, reporting of, B5.1

References useful to test user, B4.3
Regression information, I4.11
cross-validation, E10.1.1
Reliability
coefficients, F1.2, F6.2
corrected for range, F2.2
of the criterion, E4.4ff, Fff
description of procedures and samples, F2ff
of differences between scores, F6ff
express of, F3
evidence of, B5
information on test speededness, F5.2
item-test correlations, F2.4
across testers, H4
types, F
alternative forms, F4
intercorrelation of parts, F3
split-half, F5.4
test-retest, F6.1
user interpretation, J3
see also Internal consistency, Measurement error, and Stability research on experimental tests, B2.1
Revision, A3
short form, A3.2.1, A3.2.2

Sampling error, see Measurement error
Sampling methods, D2.1.2
adequacy of, in determining content validity, E *see also* Subjects
Scales, D5, D6
Scores, use of, G
Scoring of tests
basis for, C3.2
instructions, C3ff, I3.1ff
methods for checking, C3.2.1, C3.2.2, I3.1ff, J7.3
error, C3.2
system used, C3.2.1, C3.2.2, C3.2.3
Security, I5ff

Self-interpreting tests, B4.1
Short form, A3.2.1, A3.2.2
validity, A3.2.2
Significance
degree, B5.3
difference between two scores, F1.4, F6.3
validity studies, E13.2.6
Speeded tests, E13.2.4
Stability, intro. to, Fff
scores over time, F6, F6.3ff
time between measures, F6
see also Reliability
Standard error of measurement, D5.2.2, F1, F1.2
Standard scores, D3, D5.2ff, E8.1.1
Standard deviation, see Variance
Statistical techniques, E8.1.3
Subgroups
validity, E13.2.3, E13.2.5
Subjects
comparability of, E6.1.1
comparison with reference groups, E6.1.2
description, E6.2ff
experience between testings, E7.4.4
Subscores
interpretation, B5.5, E1.2

Terminology in reporting statistical procedures, F3
Test, short form of, A3.2ff

Users, types of, G
alternative interpretation, J7ff
assessment of variables, H2ff, H3ff
costs, H2.2
experience, H2.1
formulation of goals, H1ff
individual interpretation J5.3ff
institutional use, G5, I2.1
interpretation of test, Jff
method of choice of tests, H
qualifications of, G1ff, G2ff, G3ff, G4
reporting test scores, J2
validity, E2ff
use of test, H5

Validity, Eff
construct intro. to, E, E13
theoretical, E13ff, *see also* Manual, E12.1
content validity
item-course of study, E12.3
item-selection, E12.1, E12.1.1
job selection, E12.4
coefficient of, E5.2.1, E8.1.1, E8.2, E10, I4.1
group comparison, E8.3.2
combination factors, E8.5

criterion-related intro. to, E, E3, E4, E4.4, E5.1, E8
general principles, E, E8.4
local studies, E5.2.2
predictive vs. concurrent, E7.4 2, E7.4.3
occupational E3.1, E3.1.1
revision of validity claims, E6ff
content, intro. to E, E1.2ff
corrected vs. uncorrected E8.2.1
proof of, B5
psychiatric agreement, E4.4.2, E4.2.3
report of, in manual, E3.1, E5ff, E7.2, E8
types, intro. to, E1

Validity sample, E6, E12
collection of data, E7, E12
cross-validation E10ff
description, E6.2ff, E8.3ff
interpretation of E1.1ff, E1.3, E1.3.1
other relevant attributes, E12.1ff
subsamples, E1

Variance
for actual and corrected samples, E12.1
for scores, E7.4.1, E13.2
for subgroups, E9, E10.1.1, E13.2.3, E13.2.5
in validity studies, E7.4ff

Ex 4

UNITED STATES DISTRICT COURT

SOUTHERN DISTRICT OF NEW YORK

- - - - - - - - - - - - - - - - - - -x

THE GUARDIANS ASSOCIATION OF THE NEW YORK CITY POLICE DEPARTMENT, INC., THE HISPANIC SOCIETY OF THE NEW YORK CITY POLICE DEPARTMENT, INC., et al.,	:	Civil Action No. 72 Civ. 928 (S.J.R.)
Plaintiffs,	:	
-against-	:	AFFIDAVIT
CIVIL SERVICE COMMISSION OF THE CITY OF NEW YORK, et al.,	:	
Defendants.	:	

- - - - - - - - - - - - - - - - - - -x

STATE OF NEW YORK)
: ss.:
COUNTY OF NEW YORK)

RICHARD S. BARRETT, being duly sworn, deposes and says:

1. I make this affidavit in support of plaintiffs' motion for a preliminary injunction. I am presently Visiting Professor of Management Science at Stevens Institute of Technology in New Jersey, and a consultant in connection with industrial psychological problems, particularly matters dealing with fair employment issues. In this connection I am now consulting with the Pennsylvania State Police Department to make their selection and promotion procedures valid and fair for minority candidates. In the past I have consulted on employment testing with, among others, National League for Nursing, the United States Air Force, Standard Oil (New Jersey), Lever Brothers, and Equitable Life Assurance Society. I was the principal investigator on a study sponsored by the Ford Foundation which resulted in publication of the book Testing and Fair Employment, by Kirkpatrick,

UNITED STATES DISTRICT COURT

SOUTHERN DISTRICT OF NEW YORK

- - - - - - - - - - - - - - - - - - -x

THE GUARDIANS ASSOCIATION OF THE NEW YORK CITY POLICE DEPARTMENT, INC., THE HISPANIC SOCIETY OF THE NEW YORK CITY POLICE DEPARTMENT, INC., et al.,	:	Civil Action No. 72 Civ. 928 (S.J.R.)
Plaintiffs,	:	
-against-	:	AFFIDAVIT
CIVIL SERVICE COMMISSION OF THE CITY OF NEW YORK, et al.,	:	
Defendants.	:	

- - - - - - - - - - - - - - - - - - -x

STATE OF NEW YORK)
: ss.:
COUNTY OF NEW YORK)

RICHARD S. BARRETT, being duly sworn, deposes and says:

1. I make this affidavit in support of plaintiffs' motion for a preliminary injunction. I am presently Visiting Professor of Management Science at Stevens Institute of Technology in New Jersey, and a consultant in connection with industrial psychological problems, particularly matters dealing with fair employment issues. In this connection I am now consulting with the Pennsylvania State Police Department to make their selection and promotion procedures valid and fair for minority candidates. In the past I have consulted on employment testing with, among others, National League for Nursing, the United States Air Force, Standard Oil (New Jersey), Lever Brothers, and Equitable Life Assurance Society. I was the principal investigator on a study sponsored by the Ford Foundation which resulted in publication of the book Testing and Fair Employment, by Kirkpatrick,

J., Ewen, R., Barrett, R., and Katzell, R. I received a B.S. in Administrative Engineering from Cornell University in 1948, an M.A. in Education from Syracuse in 1952, and a Ph.D. in Industrial Psychology from Western Reserve University in 1956.

2. In preparing this affidavit I have examined numerous documents provided to me by plaintiffs' counsel including entry level and promotional examinations of the New York City Police Department (the "NYCPD") and the corresponding scoring keys, notices of examinations, test plans, item analyses and analyses of protests submitted by candidates taking these examinations prepared by the Department of Personnel. I also reviewed all documents submitted by defendants in response to plaintiffs' demands to produce all job analyses, job descriptions, performance evaluation criteria and validity studies in connection with the entry level and promotional examinations. The examinations to which I gave the closest attention are listed below:

a. Patrolman (Police Trainee, Policewoman) examinations 0013 and 9080,

b. Sergeant examinations 8528 and 1453,

c. Lieutenant examinations 9925, 9580 and 6512,

d. Captain examinations 1454 and 1506.

I also examined the following reports submitted to the Department of Personnel:

(1) Procedures used in New York City for Promoting Police Officers to Sergeant, Lieutenant and Captain, Educational Testing Service, June 1970, ("ETS Study, Phase I"), (2) Development of Improved Examination Procedures for Promotion of Police Officers to Sergeant, Lieutenant, and Captain in the New York City Police Department, Phase II,

Educational Testing Service, December 1970, ("ETS Study, Phase II"), (3) Development of Improved Examination Procedures for the Promotion of Police Officers in New York City, Phase III, Educational Testing Service, June 1972, ("ETS Study, Phase III"), (4) Report on Analysis of the Police Trainee Examination in Relation to Ethnic Data, Raymond A. Katzell, (5) Interim Report on Research on Police Trainee Exam # 7046, (6) Research Study on the Examination for Police Trainee, # 7046, (7) A General Critique of Exam No. 7046, Drs. Banks and Brown, (8) Department of Personnel memorandum dated January 16, 1973 concerning examinations 7046 and 9080-9081 , and (9) memoranda and correspondence by and between the Department of Personnel and the NYCPD relating to the examination process. I have also read the transcript of the deposition of Mr. Solomon Wiener, Assistant Personnel Director for Examinations for the New York City Department of Personnel, taken by counsel for plaintiffs on December 19, 1972 and February 8, 1973.

3. In evaluating the materials referred to in the previous paragraph, I have compared the procedures used by the Department of Personnel with currently accepted professional test construction practice as embodied in the Guidelines on Employee Selection Procedures of the Equal Employment Opportunity Commission, and the Standards for Education and Psychological Tests and Manuals of the American Psychological Association, to which the Guidelines refer as a primary source of professional standards in the field. Copies of those documents are annexed as Exhibits I and II hereto.

4. In my opinion, the procedures used by the

Department of Personnel to prepare, administer and grade and act upon examinations which I have reviewed and to otherwise test candidates for the positions of patrolman, sergeant, lieutenant and captain in the NYCPD do not meet accepted standards of sound professional practice in the selection testing field.

5. The first step in developing a valid selection procedure is to determine the purposes for which the procedure is to be used. It is not enough to assert that the examinations are designed to improve the quality of the work force. If a meaningful and valid test is to be developed, the goals of that test must be defined in detail. In particular, the employer requesting the examination must be specific in defining the kinds of poor performance (and poor performers) which the procedures are to exclude, or the good performance (and good performers) which the examinations are to encourage. For example, the Department of Personnel in conjunction with the NYCPD could reasonably ask such questions as: is it a significant problem in the NYCPD that patrolmen or superior officers cannot master the rules which govern their jobs; or is it a significant problem that they cannot perform under stress; or is it a significant problem that they have trouble in performing certain clerical tasks. There is nothing in the material I have reviewed which sets forth the goals of the NYCPD entry level and promotional exams in anything approaching meaningful detail. There is some indication that the fundamental purpose of the present tests is to serve as an arbitrary device to reduce the number of candidates applying for the positions to fit the available openings. Such an arbitrary goal is clearly inconsistent with proper test

preparation.

6. The second essential step in the preparation of proper selection procedures is the performance of a job analysis in order to determine what the incumbent actually does on the job, the relative importance of his various functions, the kinds and levels of the attitudes, knowledges and skills that are required, the kinds of behavior that lead to success or failure on the job, and the level of performance on significant aspects of the job that is necessary for the successful maintenance of the organization's functions. The resulting job description should consider the variations which exist in the job and the possibility for advancement within the ranks. It is virtually impossible to create a valid examination unless a job analysis and a job description have been done.

7. None of the documents which I have reviewed contain an adequate job description for an entry level or promotional examination. The brief job descriptions contained in the notices of examination prepared by the Department of Personnel are not acceptable. The material in these notices is at best cursory and is of no use to the individual actually responsible for preparing the examination.

The official description of the duties and assignments of police officers set forth in the Rules and Procedures of the NYCPD or the New York City Police Department Manual, copies of which I have reviewed generally, do not constitute an adequate job description for developing tests. They are concerned primarily with procedures which define the domain in which the patrolman or patrolwoman, or officer performs, but they give little insight into what he does, why he does it, and what types and level of skills and

knowledge are required. It is clear to me from a review of Mr. Wiener's testimony on the subject that the report of the Educational Testing Service prepared for the Department of Personnel was quite right in concluding:

> "Insofar as the E.T.S. staff has been able to determine the New York City Department of Personnel has not conducted any systematic analysis of the three superior officer grades [sergeant, lieutenant and captain]; nor have the examiners made use of other detailed job descriptions, if any do, in fact, exist in the Police Department or elsewhere in the governmental structure of New York City." (ETS Study, Phase I, p. VI-4)

8. I also concur with the statement in that same report to the effect that:

> "An essential first step toward improving the quality and the content validity of the police promotional tests lies in the preparation of up-to-date specifications based on detailed job descriptions. As far as the E.T.S. staff was able to determine, such descriptions do not exist at the present time." (Ibid., p. VI-7)

Both the quoted statements appear to be true with respect to NYCPD entry level as well as promotional exams.

9. The test plans prepared by the Department of Personnel for the entry and promotion examinations provide only a listing of the general areas in which questions are to be written, and the number of questions required for each area. An adequate test plan also includes a definition of the function to be measured, and an indication of the level of difficulty to be sought. Considerable detail is required to define psychological constructs with enough accuracy so that the item writer has a clear guide to follow. I agree with the conclusion reached by the Educational Testing Service in their June 1970 report that these test plans "provide an inadequate basis for sound test development". (ETS Study, Phase I, p. VI-4)

10. The next step in the creation of a proper test--once the job analysis has been completed and studied to determine exactly what the incumbents do, and what personal characteristics are necessary to learn and perform properly on the job--is to develop a selection procedure which measures these characteristics as accurately as possible. Tests, interviews, background investigations and any other selection devices that are used should be conceived and developed as an integrated whole and be designed to provide a maximum of relevant information so that proper selections can be assured. The NYCPD entry level and promotional examinations which I have examined emphasize intellectual functioning of dubious value for the purpose of selecting qualified patrolmen and promoting qualified patrolmen, sergeants and lieutenants to higher positions. These examinations emphasize the recognition of material from published sources, knowledge of specifics from the NYCPD rules and regulations, and estoeric abilities to visualize spatial relations, to engage in highly specific types of reasoning and to perform arithmetic calculations of doubtful relevance. My examination of the Manual and Rules and Procedures shows virtually no activity of the patrolman or officers which require the ability to visualize spatial relations or perform any but the most rudimentary arithmetic operations.

11. Once an examination has been developed, it should undergo validation studies to determine its relevance to the task at hand, in this case, identifying potentially more effective patrolmen or officers. The psychological profession recognizes three types of validity, "criterion-

related validity", "content validity" and "construct validity" for determining the adequacy or validity of an examination, of which criterion-related validity is the most appropriate measure to use for the validation of entry level and promotional examinations.

Criterion-Related Validity

12. Criterion-related validity involves many technical elements but in essence a test can be said to have criterion-related validity if it is shown that those who score high on the examination perform better on the job than those who score low. The Guidelines, § 1607.4(c) state, "Evidence of a test's validity should consist of empirical data demonstrating that the test is predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which the candidates are being evaluated." Mr. Wiener concedes that no criterion-related or predictive validity study has been completed for any NYCPD entry level or promotional examination (Wiener Deposition, p. 63-64).

13. A basic step in establishing criterion-related validity is to determine just what constitutes good performance on the job. So far as I am able to determine, this step has not been taken by defendants. No fair and logical selection and promotion procedures can be established or validated in the absence of detailed criteria as to what constitutes good performance on the job.

Content Validity

14. Content validity exists when it can be shown through careful analysis of the job and the examination that the examination is a sample of work behavior required by the

job (e.g. a typing test may be content valid for a job which requires a large volume of typing), or is a direct sample of knowledge required for successful performance (e.g. a test for medical knowledge may have content validity for licensing a physician). Content validity is an inappropriate measure to judge the validity of an examination for the selection of patrolmen because all new appointees are immediately placed in a training program to impart to them the skills and knowledge required on the job. Content validity is appropriate for promotional examinations only where it can be shown that the test measures aspects of the present job which are common to, and important to, the next level job. In addition, the claim of content validity cannot be supported without a job analysis or detailed job description giving the examination preparer adequate information upon which to select questions representing a reasonable sample of the work required for the job. As previously mentioned, I have seen no such job description or detailed job analysis done in connection with the preparation of any of the NYCPD entry level or promotional examinations.

15. On the basis of the documents I have reviewed it cannot properly be said that a study of content validity has been performed by either the NYCPD or the Department of Personnel in connection with the development of any entry level or promotional examination.

Construct Validity

16. Construct validity exists when it can be shown that the test is an adequate measure of some psychological construct, such as spatial visualization, vocabulary, verbal reasoning, or arithmetic and that the construct is logically related to job performance. To show that a test has construct validity it is necessary to define the construct,

build the measure, show that the new measure is empirically related to other measures of the same construct, and show that in other ways the test is psychometrically defensible. The documents from the Department of Personnel show that no study involving construct validity has been done, but even if they had, construct validity would be of no value unless it could also be shown that the construct was required, in the degree measured by the test, for performance on the job. As the APA Standards point out in reference to construct validity, "It must be remembered, however, that, without a study or criterion-related validity, tests developed for diagnosis or prediction can be regarded only as experimental." Since no criterion-related validity study has been done even if the examinations were shown to have construct validity they would not have been properly validated.

17. The Guidelines further require that separate validity studies be conducted for minority groups to ascertain whether the examinations are accurate predictors of future performance for each group--i.e., whether they have "differential validity." No such study has been done by the Department of Personnel or the NYCPD to determine if the entry level or promotional examinations have differential validity.

18. Although criterion-related validity is the preferable measure of an examination's effectiveness, it cannot be required where it is not technically feasible. The major reason for finding criterion-related validity studies unfeasible is a lack of a large enough number of subjects to achieve statistically reliable results. In the pool of candidates for the NYCPD there are enough Black and Hispanic subjects for such a study and, given the present state of the art, it is possible to develop procedures to assure that entry level and promotional

examinations have criterion-related validity.

19. Based upon my review of the documents previously referred to and the procedures used to prepare, administer and grade the entry level and promotional examinations, I have reached the following conclusions: (1) these examinations were not prepared in accordance with accepted and professionally recognized test construction techniques and procedure, (2) none of the entry level and promotional examinations which I have reviewed have been properly validated to assure that they test for the qualities necessary to become a satisfactory patrolman, sergeant, lieutenant or captain, as the case may be, and (3) given the present state of the art, it would be possible to create valid entry level and promotional examinations.

Richard S. Barrett

Sworn to before me this

day of March 1973

UNITED STATES DISTRICT COURT

SOUTHERN DISTRICT OF NEW YORK

- - - - - - - - - - - - - - - - - - - x

THE GUARDIANS ASSOCIATION OF THE NEW YORK CITY POLICE DEPARTMENT, INC., et al.,	:	
Plaintiffs,	:	Civil Action No. 72 Civ. 928
-against-	:	
CIVIL SERVICE COMMISSION OF THE CITY OF NEW YORK, et al.,	:	PLAINTIFFS' FIRST REQUEST FOR PRODUCTION OF DOCUMENTS
Defendants.	:	

- - - - - - - - - - - - - - - - - - - x

Pursuant to Rule 34 of the Federal Rules of Civil Procedure, plaintiffs hereby request that each of the defendants in this action produce and permit plaintiffs and their representatives to inspect and copy the documents specified below which are within their possession, custody or control or in that of any of their affiliates or representatives.

The documents requested are to be produced at the offices of Davis Polk & Wardwell, 1 Chase Manhattan Plaza, New York, New York 10005, within 30 days after the date of this request. If production of any document, or any part of a document is objected to on grounds of privilege or otherwise, each such document should be specifically identified, with a description of the document (e.g., Letter, Memorandum, Report, etc.), the general topic, the title and date, the number of pages, the name and position of the author and of each addressee or recipient, the location of each copy of such document and the grounds for refusing to produce such document.

This request is to be interpreted in accordance with the following definitions:

A. "NYCPD" refers to the defendant New York City Police Department.

B. Civil Service Commission ("CSC") means the Civil Service Commission of the City of New York.

C. Department of Personnel means the Department of Personnel of the City of New York.

D. "Entry Level Examinations" means those written examinations given to applicants for the position of patrolman or patrolwoman in the NYCPD.

E. "Promotional Examinations" means those written examinations given to applicants for the position of sergeant, lieutenant and captain in the NYCPD.

F. "NYCPD applicant screening unit" means that NYCPD administrative body which is responsible for screening entry level applicants who have passed an entry level examination for general character and fitness.

G. "Eligible lists" refers to lists promulgated by the NYCPD of persons eligible for appointment to positions of patrolman, patrolwoman, sergeant, lieutenant or captain in the NYCPD.

H. "Race-Ethnicity" refers to the characterization of described persons as black, Hispanic or white.

I. "Studies" refers not only to those studies which have been completed, but also to those which are in progress or are proposed for the future.

J. "Document" has the same meaning herein as in Rule 34(a), Federal Rules of Civil Procedure, and includes, without limitation, the original of all writings, graphs, charts, photographs, computer print-outs and other data, whether recorded, taped or otherwise coded. In any case where duplicate copies of the same document exist, all copies should be produced which contain any writing or notes which do not appear on all other copies of that document. In each case where a document is called for, all drafts of the document are also called for.

Documents to be Produced

1. Produce all documents created during the period January 1, 1965 to date which refer or relate to or which set forth:

(a) The dates of all the entry level and promotional examinations previously given or presently scheduled to be given in the future;

(b) Notices of examinations for all entry level and promotional examinations;

(c) The procedures and requirements in effect since 1961 to be followed or met by applicants for the position of NYCPD (i) patrolman or patrolwoman, (ii) sergeant, (iii) lieutenant, and (iv) captain;

(d) The number of individuals who (i) applied for, (ii) took, (iii) passed, (iv) failed and (v) withdrew

from each NYCPD entry level and promotional examination given since January 1, 1965;

(e) The name, address, phone number, race-ethnicity, examination date(s) of those individuals who applied for

(1) patrolman examination numbers 8046, 8108, 9019, 9049, 9080 and 0013

(2) sergeant examination numbers 1453, and 8528

(3) lieutenant examination number 9580

(4) captain examination number 1506;

(f) The name and examination score(s) of every applicant who took an entry level or promotional examination referred to in item 1(e);

(g) The passing score for each examination referred to in item 1(e);

(h) The lowest score received on each NYCPD entry level or promotional examination referred to in item 1(e) by an applicant who has been appointed to the position applied for;

(i) The number and race-ethnicity of individuals who have (i) taken, (ii) passed and (iii) failed each of the physical and medical examinations corresponding to each of the entry level and promotional examinations referred to in item 1(e).

2. Produce all documents created during the period January 1, 1965 to date which refer or relate to or which set forth:

(a) Eligible lists for NYCPD patrolman, patrolwoman, sergeant, lieutenant or captain in effect since January 1, 1965 as of the time each list was initially published or subsequently amended, including eligible lists

for NYCPD patrolman, patrolwoman, sergeant, lietenant and captain currently in effect from which appointments may still be made;

(b) The date when each eligible list for the position of patrolman, patrolwoman, sergeant lieutenant or captain expired and the names of those remaining on such eligible list when it expired.

3. Produce the NYCPD Personnel Research Unit, Report of December 1, 1971 giving statistics for the 1966 and 1969 NYCPD sergeant's examinations and any other document containing similar statistics for any NYCPD entry level or promotional examination.

4. Produce copies of each NYCPD entry level and promotional examination given since January 1, 1965.

5. Produce the "keys" setting forth the correct answers to each of the questions contained on each of the entry level or promotional examinations given since January 1, 1965 and all revisions of such keys.

6. Produce all documents created during the period January 1, 1965 to date which refer or relate to or which set forth:

(a) The methods, procedures and techniques for creating each of the NYCPD entry level and promotional examinations and for creating such NYCPD examinations in general;

(b) The procedures, methods and techniques used to grade each question, any subpart comprising a number of questions for which a grade was prepared, and the entire examination for each entry level and promotional examination.

7. Produce all documents created during the period January 1, 1965 to date which refer or relate to or which

set forth criticism of or suggestions for or consideration of revision of the procedures, requirements or rules involved in becoming a patrolman, patrolwoman, sergeant, lieutenant or captain in the New York City Police Department or the preparation, administration or grading of any entry level or promotional examination or category of such examinations.

8. Produce all documents created during the period January 1, 1965 to date which refer or relate to or which set forth:

(a) Any job analysis, or studies of the positions of patrolman, patrolwoman, sergeant, lieutenant and captain done with a view towards the preparation, administration, grading or other evaluation of entry level or promotional examinations;

(b) Any job descriptions or job profiles, prepared by any of the defendants or by any third party at the request or direction of any defendant, for the position of patrolman, patrolwoman, sergeant, lieutenant or captain in the New York City Police Department;

(c) The methods, procedures and criteria by which the defendants judge the actual performance of patrolmen, patrolwomen, sergeants, lieutenants or captains in the New York City Police Department;

(d) Any correlation studies which compare the success of applicants who pass given entry level and promotional examinations with their performance in the position applied for;

(e) Any studies analyzing the characteristics, background and prior training which are desirable for one in the position of patrolman, patrolwoman, sergeant, lieutenant or captain;

(f) The validity of the NYCPD entry level and promotional examinations to test for qualities viewed as being desirable in the positions applied for;

(g) Any studies indicating the "content validity", "construct validity" or "predictive validity" of the entry level and promotional examinations;

(h) Any studies, recommendations or suggestions that have been commissioned or requested by or supplied to any of the defendants by third parties with reference to entry level and promotional examinations.

9. Produce copies of NYCPD "form P.A.-15" issued since January 1, 1965 and any instructions used in connection therewith.

10. Produce any studies or evaluations of form P.A.-15 or the use of said form created since January 1, 1965.

11. Produce:

(a) Any documents created since January 1, 1965 which set forth guidelines for determining what information reported on a P.A.-15 is to be regarded as a "derogatory" or negative response.

(b) Any documents created since January 1, 1965 which set forth guidelines for assessing how such "derogatories" or negative responses are to be weighed in determining whether an applicant is suited to be a patrolman or patrolwoman in the NYCPD.

12. Produce all documents which refer or relate to or which set forth the race/ethnicity screened out on the basis of their answers to form P.A.-15 or by the NYCPD applicant screening unit since January 1, 1961 or for any lesser period of time within that period.

13. Produce all documents which relate or refer to or which set forth criteria relied upon by the CSC or the Department of Personnel in screening out applicants for entry into the NYCPD.

14. Produce all documents created since January 1, 1965 which refer or relate to or which set forth:

(a) The structure or operating procedures of the NYCPD applicant screening unit;

(b) The names, total number and race-ethnicity of those individuals who constituted the NYCPD applicant screening group as of (i) July 1, 1966, (ii) July 1, 1968, (iii) July 1, 1970 and (iv) July 1, 1972.

15. Produce all documents created since January 1, 1965 which refer or relate to or which set forth:

(a) The rules, methods and procedures used by defendants to select individuals from the various eligible lists for the positions of patrolman, patrolwoman, sergeant, lieutenant or captain;

(b) Studies by any of the defendants or by third parties validating the methods by which applicants are appointed from respective eligible lists;

(c) Procedures used by the defendants to investigate the background of those who passed the entry level examinations since 1961.

16. Produce all documents created since January 1, 1965 which refer or relate to or which set forth:

(a) The methods, rules and regulations used

to rank individuals on the eligible lists for patrolman, patrolwoman, sergeant, lieutenant and captain;

(b) Studies correlating the criteria used for placing applicants on the eligible lists for patrolman, patrolwoman, sergeant, lieutenant and captain with success criteria for performance on the job.

17. Produce all documents created since January 1, 1946 which refer or relate to or which set forth:

(a) The number and race-ethnicity of those individuals eligible to take promotional examinations for each year;

(b) The number and race-ethnicity of those holding the rank of patrolman, patrolwoman, sergeant, lieutenant or captain in the New York City Police Department for each year since 1946.

18. Produce all documents created since 1961 which refer or relate to or set forth any studies, reports, statistical data or other writings discussing the possible or actual discriminatory impact against blacks and Hispanics of any NYCPD or CSC or Personnel Department hiring and promotional practices and requirements including but not limited to the following:

(a) The requirements for taking the promotional and entry level examinations;

(b) The entry level and promotional examinations;

(c) The NYCPD physical and medical examinations;

(d) The form P.A.-15 and the questions contained therein;

(e) The practices of the NYCPD personnel screening unit;

(f) The procedures for placing individuals on the eligible lists and selecting them for appointment.

19. Produce all documents which refer or relate to or which set forth:

(a) The number and race-ethnicity of patrolmen residing in New York City at the time of their appointment for any year since 1946;

(b) The number and race-ethnicity of patrolmen residing in Nassau, Westchester, Suffolk, Orange, Rockland or Putnam Counties for any year since 1946.

20. Produce, for each NYCPD entry level and promotional examination given since 1961, any documents indicating the date when the examination was initially announced and the date when the last applicant from the corresponding eligible list was appointed.

21. Produce all documents referring or relating to consideration of giving the entry level and promotional examinations in Spanish.

22. Produce all documents referring to or relating to or setting forth the names, addresses, race-ethnicity, examination scores and passing rates for blacks, whites and Hispanics taking entry level examinations number 9080 and 9081.

23. Produce all documents which refer to or relate to or which set forth any proposals or requests for studies to be undertaken in the future concerning any of the areas covered by requests 1 to 22.

24. Produce all documents which refer or relate to or which set forth any proposals or requests for studies to be undertaken concerning new police recruiting and applicant screening procedures for the entry level examinations.

Christopher Crowley
An Attorney for Plaintiffs

JACK GREENBERG
ELIZABETH B. DUBOIS
10 Columbus Circle
New York, New York 10019
Telephone: 586-8397

CHRISTOPHER CROWLEY
1 Chase Manhattan Plaza
New York, New York 10005
Telephone: HAnover 2-3400

Attorneys for Plaintiffs

Civil Action No. 72 Civ. 928

UNITED STATES DISTRICT COURT

SOUTHERN DISTRICT OF NEW YORK

THE GUARDIANS ASSOCIATION OF THE NEW YORK CITY POLICE DEPARTMENT, INC., et al.,

Plaintiffs,

-against-

CIVIL SERVICE COMMISSION OF THE CITY OF NEW YORK, et al.,

Defendants.

PLAINTIFFS' FIRST REQUEST FOR PRODUCTION OF DOCUMENTS

JACK GREENBERG
ELIZABETH B. DUBOIS
10 Columbus Circle
New York, New York 10019

CHRISTOPHER CROWLEY
1 Chase Manhattan Plaza
New York, New York 10005

Attorneys for Plaintiffs

RECEIVED
AUG 21 1972
D.P. & W. FILES

Written
Ex 6
Ep. 13
8 11

DEPARTMENT OF PERSONNEL
CITY CIVIL SERVICE COMMISSION

NOTICE OF EXAMINATION

AMENDED NOTICE
(September 24, 1969)

No. 9080
POLICE TRAINEE (P.D.)
PATROLMAN

This examination is open only to men.

A single list will be established from this examination and appointments will be made to either Patrolman or Police Trainee (Police Department) depending on age.

Police Trainee is a trainee class of positions. A Police Trainee will receive a regular appointment as a Probationary Patrolman on reaching his 21st birthday, or as shortly thereafter as practicable, without taking any further written or physical tests, provided he has a satisfactory record as trainee and provided he passes a medical test identical to the one given to Patrolman candidates.

Salary, Vacancies and Benefits (Patrolman): The entrance salary is $9,499 per annum at present. Patrolmen receive increments of $221 at the end of their first year, $551 at the end of their second year and $679 at the end of their third year, in the Police Department. The salary reached through these increments is $10,950 per annum. Other benefits include an annual uniform allowance, 11 paid holidays annually, a generous annual leave, sick leave, membership in a liberal pension system, the Social Security System, a welfare fund, and a health insurance plan. There are many vacancies.

Police Trainee (Police Department): The entrance salary is $4,000 per annum. There will be an increment of $240 per annum after each year of trainee service up to a maximum of four increments. There will be a considerable number of positions in the Police Department under this program.

Many benefits are, or may be enjoyed by New York City employees. These benefits include generous annual leave, sick leave, membership in a liberal pension system, the Social Security System, a health insurance plan wholly paid for by The City of New York and the blood credit program.

Applications: *Filing Period* — Pre-application forms issued and received Monday through Friday from 9 a. m. to 5 p. m., except Thursday from 8:30 a. m. to 5:30 p. m., and Saturday from 9 a. m. to 12 noon from September 4, 1969 through *October 10, 1969.*

Pre-application forms are obtainable free either by the applicant in person or by his representative at the Application Section of the Department of Personnel at 49 Thomas Street, New York, N. Y. 10013. They will also be mailed on request provided that the request to the above section and address is accompanied by a stamped (6 cents) self-addressed 9½-inch envelope for each application requested. All such mail requests must be postmarked not later than *October 3, 1969.* A pre-application form submitted by mail must be received no later than the last day for receipt of pre-application forms.

Pre-applications forms are also issued in person (but not received) at all branches of the public library systems in the various boroughs of New York City, Mount Vernon, New Rochelle, White Plains and Yonkers.

Regular applications will be issued and accepted for filing in person only at the time and place of the written test.

Date and Place of Test: The written test is expected to be held October 25, 1969. This date is tentative only and may be changed if circumstances so demand.

Promotion Opportunities: The Administrative Code provides that Sergeants shall be selected from among Patrolmen of the First Grade, Sergeants are eligible for promotion examination to Lieutenant; Lieutenants are eligible for promotion examination to Captain.

Time in service as a Police Trainee shall not be considered as service in the uniformed force for pension or retirement purposes, not in computing seniority in a promotion examination, nor in granting advancement in grade as a Patrolman.

Age Requirements (Patrolman): The Administrative Code provides that only persons shall be appointed Patrolman who shall be on the date of the written test less than 29 years of age; and the Public Officers Law provides that persons shall be at least 21 years of age at the time of appointment.

Exception: All persons who were engaged in military duty, as defined in Section 243 of the Military Law, subsequent to July 1, 1949, may deduct the length of time, not exceeding a total of five years, which they spent in such military duty from the period one is allowed to be under the age limit.

Police Trainee: Candidates must be at least 18 years of age on the date of the written test and at least 17 years of age at time of appointment.

At the time of investigation, applicants will be required to submit proof of date of birth by transcript of record of the Bureau of Vital Statistics or other satisfactory evidence. Any wilful misstatement may be cause for disqualification.

Minimum Requirements: At the time of appointment: (a) graduation from a four year senior high school course, or (b) possession of an accepted high school equivalency diploma or U.S. Armed Forces Institute G.E.D. certificate (a score of at least 35 on each of the five tests and a total score of at least 225 in the examination for the diploma or certificate; or (c) certification of having passed the above examination with these scores.

At the time of appointment applicants must be United States citizens.

Applicants for patrolman must not be less than 5 feet 7 inches (three feet) in height and must approximate normal weight for height. For appointment as Police Trainee, eligibles must meet this height requirement prior to appointment. Required vision 20/30 in each eye separately, without glasses.

At the time of appointment as a Probationary Patrolman, residence in New York City, or in Nassau, Westchester, Suffolk, Orange, Rockland, or Putnam Counties is required. Also at the time of appointment as a Probationary Patrolman, possession of a valid New York State motor vehicle operator's license is required.

Proof of good character will be an absolute prerequisite to appointment. The following are among the factors which would ordinarily be cause for disqualification: (a) conviction of an offense, the nature of which indicates lack of good moral character or disposition toward violence or disorder; (b) repeated conviction of an offense, where such convictions indicate a disrespect for the law; (c) repeated discharge from employment where such discharges indicate poor performance or inability to adjust to discipline; (d) addiction to narcotics or excessive use of alcoholic beverages; (e) discharge from the Armed Forces other than the standard honorable discharge. In accordance with the provisions of the Administrative Code, persons convicted of a felony are not eligible for positions in the uniformed forces of the Police Department. In addition, the rules of the City Civil Service Commission provide that no person convicted of petit larceny, or who has been dishonorably discharged by the Armed Forces shall be examined, certified or appointed as a patrolman.

The Police Commissioner has the right to appoint one out of three on each certification.

The rules of the City Civil Service Commission provide that no name shall be certified more than three times to the same appointing officer for the same or similar position, unless at such officer's request.

Those appointed as Probationary Patrolman must serve a probationary period of one year.

A Police Trainee will not be considered a member of the uniformed force or a peace officer.

Duties of Patrolman: To perform general police duties in the various branches of the department; to perform all additional functions for the rank prescribed by relevant laws, Rules and Procedures, order or directives of the Police Department; and to perform special duties or assignments as directed by the Police Commissioner in his discretion.

Duties of Police Trainee: Following orientation and training at the Police Academy, performs, under direct supervision, routine non-law enforcement tasks in Police Headquarters or field units; performs related work.

Examples of Typical Tasks (Police Trainee): Assists precinct personnel in the preparation of records, documents, correspondence, reports, filing, typing, etc., assists precinct detectives in a clerical capacity; assists members of specialized units in their clerical tasks; operates switchboards.

Tests: Written, weight 100, 75% required.

The written test will be of the multiple choice type and will be designed to test the candidates' intelligence, initiative, judgment and capacity to learn the work of a police trainee and a patrolman. It may include questions on police situations, reading comprehension, arithmetic reasoning and civic awareness.

Candidates who pass the written test will be required to pass a qualifying physical test. This qualifying physical test may start on November 17, 1969. No second opportunity will be given to candidates who fail the qualifying physical test; or who fail to appear for the qualifying physical test, except those on active military duty.

The physical test will be designed to test the strength, agility, and power of candidate in addition, the physical test for patrolman will include a test for height. Candidates will take the physical tests at their own risk of injury, although every effort will be made to safeguard them. Medical evidence to allow participation in the physical test may be required and the Department of Personnel reserves the right to exclude from the physical test any candidate who upon examination of such evidence is apparently medically unfit. Candidates will be required to achieve a minimum mark of 70% or higher on the physical test in order to qualify. Medical and Physical requirements as posted on the Department of Personnel's Bulletin Board must be met.

Candidates shall be rejected for any deficiency abnormality or disease that tends to impair health or fitness. Such causes for rejection include but are not limited to defective vision, heart and lung diseases, hernia, paralysis and defective hearing. A history of any psychoneurotic disorder may disqualify. Persons must be free from such physical or personal abnormalities or deformities as to speech and appearance as would handicap them in the performance of their duties as a Patrolman.

Candidates who fail to attain the pass mark set for any test, subject or part of the examination shall be deemed to have failed the examination [illegible] subject, or part of the examination [illegible] to make full and complete statements on their application blanks and medical questionnaires. Misrepresentation is ground for disqualification.

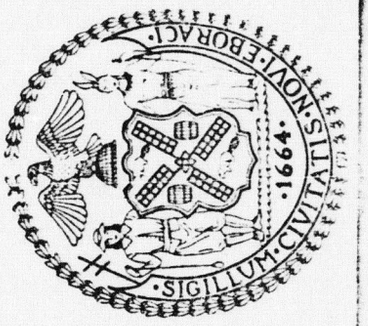
The pertinent sections of the General Examination Regulations are also to be considered part of this notice.

DEPARTMENT OF PERSONNEL, SOLOMON HUBERMAN, Personnel Director; GEORGE HENRY, Jr., Secretary. 824,26

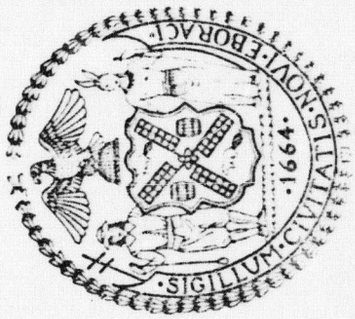
BEST COPY AVAILABLE

Bernard Ou
6.26.45

Patrolman #9049
Apl. #79653
Score 78.000
List #2414



Jaime Felix
Police Trnee P.D. # 9080
Patrolman
Apl. # 57735
score 11 - failure Code
TD. 10.25.69



THE CITY OF NEW YORK
DEPARTMENT OF PERSONNEL

NOV 21 76

NOTICE OF RESULT 7272 #

220 CHURCH STREET
NEW YORK, N.Y. 10013

0.25

7065
EXAM. NO.

TITLE: °/c Patrolman - Police Trainee NYCPD

THE SCORES YOU RECEIVED IN THIS EXAMINATION ARE SHOWN BELOW

PART	TEST SCORE	WEIGHT	PART	TEST SCORE	WEIGHT	PART	TEST SCORE	WEIGHT	PART	TEST SCORE	WEIGHT

11
FAILURE CODE

John Daley
653 Britton St, Apt A-6
Bronx, N.Y. 10467

ADJUSTED FINAL AVERAGE

LIST NUMBER

67639
APPLICATION NUMBER

VETERANS PREFERENCE

KEEP THIS NOTICE AND THE POST MARKED ENVELOPE INTACT.
SEE OTHER SIDE

INFORMATION

IF YOU FAILED THE EXAMINATION-

"FAILURE" CODE: 11 YOU DID NOT OBTAIN THE MINIMUM RATINGS REQUIRED FOR PLACEMENT ON THE ELIGIBLE LIST.
12 YOU DID NOT RANK HIGH ENOUGH TO BE PLACED ON THE ELIGIBLE LIST.

IF YOU PASSED THE EXAMINATION-

DISABLED VETERANS PREFERENCE AND VETERANS PREFERENCE—A "D" IN THE VETERANS PREFERENCE BOX INDICATES THAT TEN POINTS, IF AN OPEN COMPETITIVE EXAM, OR FIVE POINTS, IF A PROMOTION EXAM, HAVE BEEN ADDED TO YOUR FINAL EARNED RATING AND ARE INCLUDED IN THE ADJUSTED FINAL AVERAGE.

A "V" IN THE VETERANS PREFERENCE BOX INDICATES THAT FIVE POINTS, IF AN OPEN COMPETITIVE EXAM, OR TWO AND A HALF POINTS, IF A PROMOTION EXAM, HAVE BEEN ADDED TO YOUR FINAL EARNED RATING AND ARE INCLUDED IN THE ADJUSTED FINAL AVERAGE.

"PART" CODE:

- E WRITTEN TEST—PART 1
- G WRITTEN TEST—PART 2
- H WRITTEN TEST—(FINAL AVERAGE)
- J PRACTICAL TEST
- K ORAL TEST
- L TRAINING &/OR EXPERIENCE RATING
- M PRACTICAL EXPERIENCE OR TECHNICAL/ORAL TEST
- N SENIORITY RATING
- S PERFORMANCE RATING
- T PERFORMANCE & SENIORITY RATING
- U COMPETITIVE PHYSICAL TEST

LIFE OF THE LIST—THIS LIST WILL REMAIN IN EXISTENCE FOR AT LEAST ONE YEAR AND FOR NOT MORE THAN FOUR YEARS. THE MAXIMUM PERIOD ALLOWED BY THE CIVIL SERVICE LAW.

APPEALS—A CLAIM OF MANIFEST ERROR OR MISTAKE IN THE RATING OF AN EXAMINATION MUST BE MADE IN WRITING WITHIN TWO MONTHS OF THE POSTING OF THIS NOTICE OF RESULT. IF THIS IS A TRAINING AND EXPERIENCE EXAMINATION, YOU MAY SUPPLEMENT THE INFORMATION PRESENTED BY YOU ON YOUR EXPERIENCE PAPER.

NOTICE AND WARNING—THE INVESTIGATION OF YOUR CHARACTER AND FITNESS TO HOLD THE POSITION APPLIED FOR WILL BE DETERMINED AFTER YOU FILE ON NOTICE FROM THIS DEPARTMENT, A PERSONAL HISTORY SHEET. ALL STATEMENTS MADE THEREIN AND ALL STATEMENTS MADE IN ANY EXPERIENCE QUESTIONNAIRE PREVIOUSLY FILED WILL BE THOROUGHLY INVESTIGATED. ANY FALSE STATEMENTS MADE BY AN ELIGIBLE MAY CAUSE HIS DISQUALIFICATION. IMPERSONATION IN ANY EXAMINATION IS A MISDEMEANOR AND WILL BE PROSECUTED AS SUCH. ELIGIBLES ON THIS LIST MUST NOTIFY THIS DEPARTMENT PROMPTLY OF ALL CHANGES OF ADDRESS. WHEN WRITING TO THIS DEPARTMENT ALWAYS MENTION THE CORRECT LIST TITLE, ADJUSTED FINAL AVERAGE AND ORIGINAL NUMBER ON THE LIST.

RETAIN THIS NOTICE FOR YOUR FULL PERIOD OF ELIGIBILITY—ALWAYS BRING THIS NOTICE WITH YOU WHEN REQUESTING INFORMATION.

PARTICIPATION OF TOTAL UNIFORMED PERSONNEL
IN NEW YORK CITY POLICE DEPARTMENT
BY RACE AND SEX

AS OF DATE: 4/3/75

Prepared by the Equal Opportunity Unit

RANKS		TOTAL IN RANK	MALE: TOTAL MALE IN RANK	MALE: WHITE	MALE: BLACK	MALE: HISPANIC	FEMALE: TOTAL FEMALE IN RANK	FEMALE: WHITE	FEMALE: BLACK	FEMALE: HISPANIC
Chief of Oper.	#	1	1	1						
	%	100	100	100						
Bureau Chiefs	#	5	5	4	1					
	%	100	100	80.0	20.0					
Assist. Chief	#	15	15	14		1				
	%	100	100	93.3		6.7				
Deputy Chief	#	23	23	22	1					
	%	100	100	95.7	4.3					
Inspector	#	43	42	40	2		1	1		
	%	100	97.7	93.0	4.7		2.3	2.3		
Dep. Inspector	#	102	102	99	3					
	%	100	100	97.1	2.9					
Captain	#	356	357	354	2	1	1	1		
	%	100	99.7	98.9	0.5	0.3	0.3	0.3		
Lt. Dir. Legal Bur.	#	1	1	1						
	%	100	100	100						
Lt. Sp. Assign.	#	15	15	14	1					
	%	100	100	93.3	6.7					
Lt. Co. Det. Sq.	#	57	57	54	2	1				
	%	100	100	94.7	3.5	1.8				
Lieutenant	#	985	980	948	25	7	5	5		
	%	100	99.5	96.3	2.5	0.7	0.5	0.5		
Sgt. Special Assign.	#	23	23	21	2					
	%	100	100	91.3	8.7					
Sgt. Sup. Det. Sq.	#	89	89	82	5	2				
	%	100	100	92.1	5.6	2.3				
Sergeant	#	2,731	2,716	2,585	90	41	15	14	1	
	%	100	99.4	94.6	3.3	1.5	0.6	0.5	0.1	
Detective First	#	131	128	110	14	4	3	2	1	
	%	100	97.7	84.0	10.7	3.0	2.3	1.5	0.7	
Detective Secon.	#	581	574	487	65	22	7	6	1	
	%	100	98.8	83.8	11.2	3.8	1.2	1.0	0.2	
Detective Third & Sp.	#	2,946	2,895	2,449	313	133	51	39	12	
	%	100	98.3	83.2	10.6	4.5	1.7	1.3	0.4	
Police Officer Sp. Assign	#	35	35	33	1	1				
	%	100	100	94.2	2.9	2.9				
Police Officer	#	22,581	21,942	19,369	1,787	786	639	465	159	15
	%	100	97.2	85.8	7.9	3.5	2.8	2.0	0.7	0.1
TOTAL	#	30,722	30,000	26,687	2,314	999	722	533	174	15
	%	100	97.6	86.9	7.5	3.2	2.4	1.7	0.6	0.1

TOTALS - MALE AND FEMALE

WHITE	BLACK	HISPANIC	TOTAL
27,220	2,488	1,014	30,722
88.6%	8.1%	3.3%	100%

9A

PARTICIPATION OF TOTAL UNIFORMED PERSONNEL
IN NEW YORK CITY POLICE DEPARTMENT
BY RACE AND SEX

As of
Date: 8-28

Prepared by the Equal Opportunity Unit

		MALE							FEMALE					
RANKS	TOTAL IN RANK	TOTAL MALE IN RANK	WHITE	BLACK	HISPANIC	ASIAN AMERICAN	AMERICAN INDIAN	OTHER	TOTAL FEMALE IN RANK	WHITE	BLACK	HISPANIC	ASIAN AMERICAN	AMERICAN INDIAN
All Captains and Above	# 529	527	516	9	2				2	2				
	% 100	99.6	97.5	1.7	0.4				0.4	0.4				
Chief of Operation	# 1	1	1											
	% 100	100	100											
Bureau Chiefs	# 5	5	4	1										
	% 100	100	80	20										
Assistant Chief	# 14	14	13		1									
	% 100	100	92.3		7.1									
Deputy Chief	# 23	23	22	1										
	% 100	100	95.7	4.3										
Inspector	# 42	41	38	3					1	1				
	% 100	97.6	90.5	7.1					2.4	2.4				
Deputy Inspector	# 98	98	96	2										
	% 100	100	98.0	2.0										
Captain	# 346	345	342	2	1				1	1				
	% 100	99.7	98.8	0.6	0.3				0.3	0.3				
All Lieutenants	1,038	1,033	998	27	8				5	5				
	% 100	99.5	96.1	2.6	0.8				0.5	0.5				
Lt. Jr. Legal B.	# 1	1	1											
	% 100	100	100											
Lt. Sp. Assign.	# 16	16	16											
	% 100	100	100											
Lt. Det. Sqd.	# 56	56	53	2	1									
	% 100	100	94.6	3.6	1.8									
Lieutenant	# 965	960	928	25	7				5	5				
	% 100	99.5	96.2	2.6	0.7				0.5	0.5				
All Sergeants	# 2,833	2,817	2,679	95	43				16	15	1			
	% 100	99.4	94.6	3.3	1.5				0.6	0.5	0.04			
Sgt. Sp. Assign.	# 24	24	22	2										
	% 100	100	91.7	8.3										
Sgt. Supv. Det. Sqd.	# 83	83	77	4	2*									
	100	100	92.8	4.8	2.4									
Sergeant	# 2,726	2,710	2,580	89	41				16	15	1			
	% 100	99.41	94.65	3.26	1.50				0.59	0.55	0.04			
All Pol. Off. & All Det.	22,995	22,704	20,117	1,837	734	6	8	2	291	224	64	3		
	% 100	98.7	87.4	8.0	3.2	0.03	0.05	0.03	1.3	1.0	0.28	.01		
Det. First	# 124	121	106	12	3				3	2	1			
	% 100	97.58	85.48	9.68	2.42				2.42	1.61	0.18			
Det. Second	# 556	548	469	58	21				8	7	1			
	% 100	98.56	84.35	10.43	3.78				1.44	1.26	0.18			
Det. Third & Sp.	# 2,917	2,866	2,420	310	134		1	1	51	39	12			
	% 100	98.25	82.89	10.63	4.59		0.10	0.03	1.75	1.34	0.41			
Police Off. Sp. Assign.	# 35	35	33	1	1									
	% 100	100	94.2	2.9	2.9									
Police Officer	# 19,363	19,134	17,089	1,456	575	6	7	1	229	176	50	3		
	% 100	98.81	88.22	7.52	2.97	0.03	0.04	0.03	1.18	0.91	0.26	0.02		
Total	# 27,395	17,081	24,310	1,968	787	6	8	2	314	246	65	3		
	% 100	98.85	88.71	7.18	2.87	0.03	0.04	0.03	1.115	0.90	0.24	0.01		

TOTALS - MALE & FEMALE

WHITE	BLACK	HISPANIC	ASIAN AMER.	AMER. INDIAN	OTHER	TOTAL
24,556	2,033	790	6	8	2	27,39
89.64%	7.42%	2.88%	0.02	0.03%	0.01%	100.00%

9B

78100-015

LAW DEPARTMENT
MUNICIPAL BUILDING, NEW YORK, N. Y. 10007

NORMAN REDLICH, *Corporation Counsel*

Mr. Allan Luks
Secretary-Treasurer
New York City Rand Institute

Dear Mr. Luks:

This letter summarizes the agreement between the Law Department and the New York City Rand Institute in regard to a study related to federal civil action #72-928 (The Guardians Association of the New York City Police Department, et al. v. Civil Service Commission of the City of New York, et al.).

1. The purpose of the study is to determine the distribution of civil service examination scores by race for those persons who took exam number 8046 and/or exam number 0013, and to perform statistical tests giving the probability that the observed differences in the distributions by race could have occurred by chance. The distributions will be calculated both for the score on the written exam and for the final score which includes veterans' preferences.

2. The study will begin on November 9, 1971 and the final results will be provided before March 13, 1973. Although a completion date of March 8, 1973 was originally agreed upon, a delay has been necessitated because the computer printouts of exam number 8046 do not contain the names, addresses or test scores of those who failed the exam, while those of exam number 0013 contain the names but not the scores of those who failed. The Department of Personnel is conducting a search to provide the necessary information.

3. The Institute will be reimbursed for the actual cost of the study plus a fixed fee of $1,868, with the total cost plus fixed fee not to exceed $33,500.

4. Funding is expected to be provided through the Fund for the City of New York. At the present time, the Fund has authorized expenditures not to exceed $10,000; if actual payment of this sum is not received by the Institute by November 15, 1972, the study will terminate until such payment is received. If authorization for the remaining $23,500 is not received by December 25, 1972:

a. the Institute will terminate the study;

b. the Institute will provide the Law Department with any and all data collected prior to that date, in unprocessed form:

c. the Institute will provide a statement of work tasks completed and expenditures incurred for each task, but will not submit a report of findings; and

d. the Institute will be reimbursed for actual costs incurred plus a fixed fee of $565, with the total not to exceed $10,000.

5. To permit this study to be conducted, the City will supply the Institute with the names, addresses, and test scores of all examinees (including veterans' preferences) for exams 8046 and 0013; access to Police Department files which contain relevant information about applicants who passed these exams; computer processing to attach a census tract number to the address record of each subject who lived in New York City; and any other data available to the City which will further the purposes of the study. None of the data received by the Institute will be disclosed to the public in a form which would permit identification of the race, exam score, or address of any individual subject.

6. The results of this study, and descriptions of the methodology used, will be made available to all parties to this civil action simultaneously, and the Institute may publish such results and descriptions under the names of the authors, with the report to contain a statement to the effect that neither the methodology nor the findings necessarily represent the views of the City, any party to this civil action, or the Fund for the City of New York.

7. During the course of this study, the Institute will mail questionnaires to the subjects and will telephone or visit the homes of some subjects. The format of the questionnaire will be approved in advance by all parties to this action and will include a request for a response signed by appropriate representatives of one or more parties to the action. Also, the City will cooperate fully so that Institute staff can conduct the proposed home visits as quickly and efficiently as feasible.

8. The principal investigators for this study will be Jan M. Chaiken and Bernard Cohen. The Institute will assign such other personnel as it deems necessary.

If this letter correctly summarizes our understanding would you kindly sign and return the attached copy.

Sincerely,

Norman Redlich

Corporation Counsel

Allan Luks

Alan Luks
Secretary-Treasurer
New York City Rand Institute

cc: Christopher Crowley, Esq.
1 Chase Manhattan Plaza
New York, N. Y. 10005

98700-010

10-26-72

THE NEW YORK CITY-RAND INSTITUTE PROPOSAL: ANALYSIS PERTAINING TO CIVIL ACTION #72-928

The essence of the New York City-Rand Institute proposal for analysis related to civil action #72-928 is as follows:

(1) Questionnaires will be mailed to all persons who took civil service exams numbered 8046 and 0013.

(2) The questionnaire will ask, among other things, the race of the recipient, with the question to be phrased in a form suitable to all parties to the action. Also the letter accompanying the questionnaire will request a member of the household of the intended recipient to indicate the race of the addressee if he is not available to return the questionnaire himself.

(3) Positive racial identification will be made of all subjects who returned their PA-15 form.

(4) Specific techniques described below will be used to identify the race of non-respondents.

(5) The distributions of scores, both before and after addition of veterans' preferences, will be calculated by race for the subjects, and statistical tests will be performed to determine the probability that the observed differences in the distributions by race could have occurred by chance.

In the absence of additional information, the distributions calculated according to this procedure must be assumed to be biased estimates of the actual distributions, since the respondents may differ in unknown ways from the non-respondents. Three tests of the "typicality" of the respondents can be made at each stage of our analysis. If all three tests are positive, this would suggest strongly that the respondents do not differ in important ways from the non-respondents. These tests, which will be performed, are:

(1) whether the overall distribution of exam scores for the respondents is not significantly different from the distribution for the non-respondents.

(2) whether the geographical distribution of responses is not different statistically from the geographical distribution of non-respondents.

(3) whether the numbers of respondents in each racial category are not significantly different from the numbers which would be expected if respondents were drawn randomly from the group of males aged 20-29 in their census tract.

It may not be reasonable to expect each of these tests to be positive. The Institute has previously mailed a questionnaire to all persons who applied for the patrolman's exam of April 5, 1969, but did not appear to take the exam. Among the respondents, 29.4 percent indicated their race as black and 14.4 percent indicated Puerto Rican. This compares with 18.1 percent blacks and 11.1 percent Puerto Ricans in the indicated age range in the general population of New York.

Therefore, appropriate steps will have to be taken to improve the response rate and/or to estimate the distribution of scores by race for the non-respondents. A 40 percent sample of the non-respondents will be selected and the following techniques can be used, either singly or in combination to establish each subject's race:

(1) Non-respondents can be sent a follow-up questionnaire.

(2) An attempt can be made to reach the addressee or a member of his household by telephone.

(3) Non-respondents whose last name is included in the Census Bureau's list of 8000 Spanish surnames can be identified as "presumably Hispanic."

(4) An attempt can be made to reach the non-respondents by sending someone to his address.

(5) The distribution of scores by race for the remaining non-respondents can be estimated by using census data giving the race of males aged 20-29 in each census tract.

Technique (1 - follow-up questionnaire) does not appear feasible due to the time constraints on this study. Technique (2 - phone calls) will be used in an attempt to contact all subjects in the sample. Census data indicate that about 14 percent of the population cannot be reached at home by telephone, with percentages ranging from 7.4 in Staten Island to 22.3 in the Bronx. Black and Hispanic persons are much more likely than whites to be unreachable by telephone; for example, in Brooklyn 36.3 percent of blacks cannot be reached by phone, compared to 16.9 percent of whites. Therefore, an attempt to determine by telephone the race of persons who did not respond is likely to leave a residual group of nonrespondents which is predominantly black and Hispanic.

The fraction of persons who will refuse to give their race on the telephone is not known to us but may not be negligible. As a rough estimate we might say that 60 percent of the non-respondents will give their race over the telephone.

Technique (3 - surnames) is feasible and will be used. If the wording of the questions on the questionnaire are suitable, the error rate for technique (3) can be estimated from data supplied by respondents. In the list of questions which follow, (a^1), (b^1), (c), and (c^1) are examples of formats suitable for comparison with the list of Spanish surnames, while (a) and (b) are suitable for comparisons with census tract data in New York.

(a) To which of the following groups do you belong:

_____ Negro or Black

_____ Puerto Rican

_____ White

_____ Other _____

(a^1) Same as (a), but replace "Puerto Rican" with "Hispanic" or "Spanish American" or "Spanish-speaking American."

(b) Were you or either of your parents born in Puerto Rico? (yes or no)

(b^1) Were you or either of your parents born in Mexico, Cuba, Puerto Rico, Central or South America?

(c) Is Spanish spoken in your home?

(c^1) Was Spanish spoken in your home when you were a child?

If time and funds are available, then home visits (technique -4) could be made to find the hard-core non-respondents. Geographical stratification of the sample will assist in this. However, a possible alternative to extensive home visits consists of estimates based on a small number of visits plus use of technique (5 - census tracts) which will yield equally reliable results more rapidly. The approach would be as follows:

(1) It is possible to determine from the Census Bureau's data tapes the number of males aged 20-29 who are Puerto Rican, black but not Puerto Rican, and white but not Puerto Rican in each census tract.

(2) From questionnaire responses, telephone follow-up, and even a small number of address visits, and matching of surnames, we expect that

for a number of census tracts the race of all persons who took the exam and lived in that census tract will be known. This will permit calculation of correction factors relating census-tract race distributions to exam-taker race distributions.

(3) From this information, plus the data available from respondents and census tapes, we can estimate the probabilities pj1, pj2, pj3 that an exam taker from census tract j is Puerto Rican, black, or other. An unbiased estimator of any parameters of the Puerto Rican, black, and "other" score distributions, s_1, s_2, and s_3 (which might be fraction passing, fraction in any score range, or mean score), can then be calculated as follows:

$$\underline{s} = C^{-1}\underline{z}$$

where C is a matrix with components

$$C_{ij} = \frac{\sum_{\ell=1}^{k} n_\ell p_\ell \; p_{\ell j}}{\sum_{m=i}^{k} n_m o_{mi}}$$

and

$$z_i = \frac{\sum_{\ell=1}^{k} n_\ell p_{\ell i} s_\ell}{\sum_{m=1}^{k} n_m o_{mi}}$$

Here n_ℓ is the total number of exam-takers in census tract ℓ, and S_2 is the average of the observed values of the parameters of interest in census tract ℓ. K is the number of census tracts.

The variance of the parameters estimated in this fashion can also be calculated as functions of n_ℓ and $p_{\ell i}$. The magnitude of the variance, which is of interest because it indicates the extent to which the estimates can be improved over what was known from the questionnaires and telephone calls alone, unfortunately cannot be calculated until

all the data are available for processing. However, considering the number of exam-takers in the group to be studied, we feel the variance will be very small compared to s_i itself.

The following example summarizes the procedures for identifying the race of approximately 10,000 persons who took the two exams in question:

Techniques of Identification	Number Remaining	Remaining in Sample
(Approximately 10,000 men took exam)	(10,000)	-
1. Approximately 2000 men were appointed from the two lists - PA-15 Identification.	8,000	-
2. An estimated 2000 men submitted PA-15s but were not appointed - PA-15 Identification.	6,000	-
3. We expect at least a 30 percent questionnaire response, (1800 subjects).*	4,200	-
4. A 40 percent sample of non-respondents will be selected (1680 men).		1,680
5. An estimated 60 percent of the subjects will be identified by phone calls (or 1008 men).		672
6. We estimate about 72 men will be identified through surnames.		600
7. Approximately 300 persons will be identified through home visits.**		300
8. The race of the 300 hard-core can be established through census tract analysis.		0

---

* In an Institute study, questionnaires sent to persons who applied for, but did not take, the 1969 exam resulted in a 35.5 percent response rate. These questionnaires permitted an anonymous response, while the questionnaire for the present study cannot permit anonymous response; therefore a lower response rate can be expected. (Part of this may be neutralized since the subjects in the present study not only applied for the exam but also took it.)

** In our Institute study 5.6 percent of the respondents could not be located by the post office.

STATISTICAL PREPARATION FOR GUARDIANS ASSOCIATION
AND HISPANIC SOCIETY VS.
THE CIVIL SERVICE COMMISSION AND THE POLICE DEPARTMENT
(PROPOSAL)

This proposed study is directed at resolving certain questions of fact raised in the complaint by the Guardians Association and the Hispanic Society against the Civil Service Commission, the Department of Personnel, and the Police Department (Civil Action No. 72-928). Only those questions which are related to the practices and procedures for selection and appointment of patrolmen are to be considered in this study.

The questions to be resolved are as follows:

(1) Among those persons who take the civil service entrance examination for patrolman, does the proportion passing differ by race in a statistically significant way (i.e., in a way which is unlikely to be explained by chance alone)?

(2) Is the proportion of black or Hispanic persons scoring in each 10-percentile range on the written civil service examination significantly different from the proportion of whites?

(3) Does the addition of veteran's preference credits to the written examination score affect the ranking of persons on the appointment list in a way which differs significantly by race?

(4) Among those who pass the civil service examination, does the proportion who appear for the medical/physical examination differ significantly by race?

(5) Among those who appear for the medical/physical examination, does the proportion who pass differ significantly by race?

(6) Among those who pass the medical/physical examination, does the proportion who complete and return their PA-15 differ significantly by race?

(7) Among those who file completed PA-15 forms, does the proportion who are accepted, rejected, or withdraw differ significantly by race?

The answers to these questions will determine whether the selection/appointment process does in fact discriminate by race and, if so, will determine which parts of the selection process make the largest contribution to the observed disparities by race. The outcomes of similar suits in other cities suggest that discrimination, if it does occur, may not be primarily due to the civil service examination. We believe it is important to study the entire process.

Note, however, that the proposed study will not determine whether the observed differences by race are warranted in light of the background, education, health status, intelligence, or estimated effectiveness as a police officer of those who appear for the examination. Moreover, the results of the study will not necessarily clarify the reasons for the disparities which are found to exist. For example, the complaint alleges:

- That the written examinations place an overemphasis on reading and verbal skills;
- That persons whose primary language is Spanish are discriminated against by an examination given in English;
- That the voluminousness and complexity of the PA-15 disproportionally discourages minority group members from completing it; and
- That the predominance of white screening officers has a chilling effect on minority applicants.

We believe that a statistical analysis of the type proposed here will be more complete, accurate, and reliable if it is restricted to questions of fact and does not address judgmental allegations such as these. If the results indicate discrimination in one or more phases of the selection process, a later study could explore the contributory factors, if desired.

Our approach to answering questions 1 to 7 posed above will be to utilize a combination of mailed questionnaires and statistical analysis of census data. We propose to consider the entire group of persons who appeared for and took civil service exams numbers 8046 (July 20, 1968) and 0013 (May 9, 1970).

It will be necessary for the Department of Personnel to supply us with a copy of its computer tapes which have the name, address at time of examination, examination score, adjusted score, and rank on the appointment list of each examinee for these two examinations. These tapes will be used to address questionnaires to be mailed to the examinees. Although the information we need about persons who completed PA-15 forms is available from the Police Department, the elimination of the names of such persons from the mailing list may take longer than the preparation of the questionnaire, in which case all examinees would be sent the questionnaire.

We propose to prepare the questionnaire on a printed, two-part postcard with an addressed, postage-paid section for return to us. The request to the examinee to return his completed questionnaire should be signed by the Personnel Director or the Corporation Counsel. The text and format of the postcard will be submitted for approval to all parties in this civil action.

The questions will ask the race of the respondent, whether he passed the civil service written examination (to determine the extent to which examinees find out the results), whether he appeared for the medical/physical, whether he passed the medical/physical, and whether he submitted a form PA-15. It is important for the parties to agree on the definitions of racial categories to be used on the questionnaire.

In addition, the race of all examinees who completed PA-15s will be determined independtly from their application folders in the Police Department, and we will also determine from PD files whether they were accepted, rejected, or withdrew. The method of collecting these data is familiar to us since we previously obtained the same information for an earlier examination. (See Hunt, I.C., and B. Cohen, Minority Recruiting in the New York City Police Department.)

To determine the desired information by race for those subjects of the study who do not respond to the mailed questionnaire, we propose to assign "probabilistic races" to each person based on the census tract of his address and the racial composition of males aged 20 to 29 in that tract. Estimates of the total number by race in each category of interest (pass/fail, etc.) will be obtained from these probabilistic races. This method will be checked for accuracy by comparing the estimates with actual data for those subjects who return their postcard questionnaire and for those who completed PA-15s. This step in the analysis will enable us to determine whether the group of respondents differs substantially from the total group. We believe the results obtained in this way will be more reliable and complete than could be achieved from follow-up questionnaires or phone calls.

To enable us to perform this anlysis, it will be necessary for the Law Department or the Department of Personnel to request use of the City's GIST/ADMATCH programs for obtaining census tracts from the addresses on the computer tapes. These programs, operated by the City Planning Department, were developed by Robert Amsterdam, Office of the Deputy City Administrator, and are made available only to city agencies.

The proposed time-table for this study, assuming the GIST programs are used, is as follows, beginning at the time a copy of the appropriate computer tapes is obtained:

Item	Time
Prepare and mail postcards	By end of week 3
Collect data from PA-15s	By end of week 5
Obtain computer tape with census tracts added	Week 6
Keypunch PA-15 data and responses	Week 7
Merge data and process data	Week 8
Preliminary analysis of responses	Week 9
Match file with census data	Week 9
Update for late responses	Week 10
Final analysis of data	By end of week 12
Written final report	By end of week 16

The principal investigators for this study will be Bernard Cohen and Jan Chaiken. The costs will be billed as actually incurred, not to exceed the following:

Personnel Costs	$16,000
Overhead	6,883
Travel	600
Computer	1,400
Printing Postage	1,500
Subtotal	$26,383
General & Administrative Charge	4,749
Total Costs	$31,132
Fixed Fee	1,868
Contract Total	$33,000

17 I

DETAILED REPORT ON RESEARCH

ON POLICE TRAINEE EXAM #7046

The traditional method of personnel selection by the City of New York's Personnel Department is through competitive written and oral tests. Several studies such as were done at Berkeley, California and at the New York Port Authority have indicated that this selection process screens out members of minority groups as they do not do well on these types of tests.

In 1967 the New York City Department of Personnel began a study to determine whether certain minority groups were, in fact, being excluded from City jobs because they did not pass our written tests. We suspected that members of these groups may have had the aptitude to successfully perform the jobs they applied for, but were prevented from doing so because of below-passing test grades which may have been unrelated to actual job performance.

Our first step was to establish whether our tests did discriminate against minority groups, and, if so, to what extent. The exam with which we chose to start the study was for Police Trainee, which was held November 4, 1967. The key advantages of using this exam were that, as an entrance-level, open-competitive one, we would be able to study a large, as well as socially and economically heterogeneous, candidate population.

A Police Trainee, after orientation and training at the Police Academy, performs, under direct supervision, routine non-law enforcement tasks in Police Headquarters or field units. On reaching his 21st birthday he

receives a regular appointment as a Patrolman, without taking any further written or physical examinations, provided he has a satisfactory record as a trainee and passes a medical test identical to the one given to Patrolman candidates.

The test given to the candidates that November consisted of 100 objective-type items involving judgment, spatial relations, verbal analogies, vocabulary, arithmetic reasoning and reading comprehension. If we found the test discriminatory, we would then analyze these subject areas, or sub-tests, to determine the extent of the discrimination.

A long-range phase of our study was to determine which test factors best predict successful performance at the Police Academy, and do not invalidly eliminate members of minority groups. Criteria for this are to be the grades and other evaluations from the candidates' records at the Police Academy.

Those items, if any, which predict good performance, but tend to be answered incorrectly by minority group candidates, were to be separated out. These questions would be compared to those items, if any, which predict good job performance but do not discriminate against minority group members. As a first step toward this end we would review the test items to ascertain if we had any items which were answered equally well by all groups. From this analysis we hoped to identify what makes a question a valid indicator of successful job performance, and to isolate those factors which discriminate against minority candidates.

Because several studies had shown that socio-economic status has some relationship to test scores, we studied the nature of that relationship.

To do so, we designed and analyzed the data supplied by a personal history questionnaire, which was answered by those candidates who were selected to be the subjects.

The procedure for administering the Police Trainee exam was the usual procedure used by the Department of Personnel. Candidates were asked to report to various high schools to take the written test at 9 A.M. They were given seating cards directing them to a particular room and seat in that room. These seating cards had been scrambled so that each candidate was randomly seated.

Of the 4079 candidates who appeared for the test, 1243 were selected by the following procedure to answer a personal history questionnaire: every third room was selected at random and all candidates in these rooms were directed, upon completing the written test, to go to a nearby room. There, they were asked to fill out the questionnaire. Although this was completely voluntary, only three candidates refused.

The following personal data were compiled from the questionnaires:

Ninety-five per cent of the candidates ranged in age between 15 and 21, with the mean at 17.7 years. Only 67 candidates were married; 3% of the white candidates were married, 13% of the Negroes, and 12% of the Puerto Ricans.

They came from the five boroughs of New York City and three surrounding counties -- Nassau, Suffolk and Westchester. However, only 21% of the whites resided in New York County, that is, Manhattan, whereas 43%

of the Negroes and 70% of the Puerto Ricans lived in New York County. Just 6% of the Puerto Ricans came from the borough of Queens. The white population was distributed throughout the area, but the Negro candidates came mostly from New York County, Kings County and Queens County, while the Puerto Ricans were mostly from New York County, Kings County and the Bronx.

The group was composed of 17% Negroes, 12% Puerto Ricans and 71% whites. English was spoken in the homes of 92%, Spanish in the homes of 14% and Italian in the homes of 11%.

Their education ranged from elementary school graduation to college graduation. The average number of years of schooling the population had beyond elementary school was 3.4 years. Forty-six per cent of the population had a high school diploma. Of these, 44 % had a general diploma, 33% an academic diploma, 11% a vocational diploma, 8% a commercial diploma, and 7% had an equivalency diploma.

Including part-time employment, 54% of the candidates reported they had jobs. Fifty-five per cent of the white candidates were employed, 54% of the Negroes, and 49% of the Puerto Ricans. Ninety-one per cent of the white group's fathers were employed, compared to 86% of the Negroes' fathers and 77% of the Puerto Ricans' fathers.

Forty-one per cent of the fathers of the white and Negro candidates were high school graduates; but only 15% of the fathers of the Puerto Rican candidates had a high school diploma, and 47% had finished only elementary school.

Forty per cent of the white group's mothers were employed, whereas 50% of the Negro candidates' mothers had jobs and 34% of the Puerto Rican candidates' mothers had jobs. Fifty-eight per cent of the white group's

mothers but only 42% of the Negroes' mothers had a high school diploma, while only 8% of the Puerto Ricans' mothers had graduated from high school. Sixty-eight per cent of the Puerto Rican candidates' mothers had gone only as far as elementary school, and none had attended or graduated from college. Of the Negro candidates' mothers, 7% had some college and 4% were college graduates; while 3% of the whites' mothers had attended college and 4% had graduated.

At a later date we sent follow-up questionnaires to candidates who had answered the personal history questionnaire and passed the written test, but failed to show up for the medical and physical examinations.

Of the 108 who replied, 34% said they were in the military service, 7% were no longer interested, 5% wished to keep their present jobs, 2% said the money was inadequate, 13% were still in school, 4% didn't show up for medical reasons, 2% were away from home at the time, 3% had been appointed from previous lists, 11% intended to apply for the next exam, and 20% had no intention of applying for the next exam.

To analyze the results of the written exam and compare them with the background data, we used a sample of 1216 men, 858 who reported they were white, 212 who were Negro and 146 who were Puerto Rican. As it turned out, this ethnic factor (that is, whether the applicant was white, Negro or Puerto Rican) emerged as the only background factor that had a marked and consistent association with test performance.

White candidates achieved the highest average scores on the total test and for each of its six sub-tests; their average score was 70.5. Puerto

Ricans had the lowest average score -- 49.1. And Negroes fell in the middle, with an average score of 57.9. Our analysis also reveals that candidates with higher socio-economic status, who were usually whites, had higher scores. Candidates who lived in more segregated areas, those with the highest concentration of Negroes, had lower scores.

We reviewed the 100 test items to determine whether any had not differentiated between Negroes and whites and between Puerto Ricans and whites. We found a number of items on which Negroes and Puerto Ricans performed within 11% of whites. These items were generally the easier test items for the minority populations. That is, of the 45 items where the Negroes scored within 11% of whites, the mean score for Negroes was 69%. However, on the remaining 55 items, where the difference between whites' and Negroes' scores was greater than 11%, the mean score for the Negroes was only 50%. The difference of 19% in the Negroes' performance on these two groups of test items is statistically significant.

In a similar vein, Puerto Ricans scored within 11% of the white population on 20 items. Their mean score for these items was 65%. On the remaining 80 items, the Puerto Rican population's mean score was only 46%. This difference of 19% was statistically significant.

We reviewed the questions to determine if there were other findings which would be relevant. We found that, by a subjective evaluation of items in the judgment sub-test, those items which were more concrete or more relevant to the minority groups' experience tended to fall into the group of 45 items where the two minority populations scored within 11% of the white population. For example, more of the minority groups knew how to solve a practical street problem than knew the reason for a particular official action. Also, we found

that the greatest disparity in performance of the white population and the Puerto Rican population was on the vocabulary and verbal analogies sub-tests. This is understandable in light of the language difficulties that Puerto Ricans probably have.

A word of caution is in order. This is an interim report. So our findings are only tentative. Since we have not yet completed our study, we don't know whether the difference in scores reflects a built-in bias or a true difference in ability to perform the duties of the job.

Director, Bureau of Examinations A

Examining Division R

April 13, 1971

Interim Report on Research on Police Trainee, Patrolman, Exam. No. 9030 and Policewoman, Exam. No. 9031

In 1967 the New York City Department of Personnel began a study to determine whether minority groups were being unfairly excluded from City jobs. This would be considered the case if we found that members of minority groups who may have had the aptitude to succeed on the job were prevented from being hired because they failed written tests which were not related to actual job performance.

The examination chosen for this study was for the position of Police Trainee, held on November 4, 1967. To determine which factors influence performance on the written test, 1243 candidates were selected at random and asked to complete personal history questionnaires. The group was composed of 17% Negroes, 12% Puerto Ricans, and 71% whites. From analysis of all available data, ethnic group membership emerged as the only background factor that had a marked and consistent relationship to test performance. Whites achieved the highest scores on the total test (mean 70.5) and on each subtest; Puerto Ricans had the lowest scores in both categories (mean on total test 49.1); and blacks scored between them (mean on total test 57.9).

The 100 test items were then reviewed to determine whether any had not differentiated between whites and Negroes or between whites and Puerto Ricans. Negroes scored within 11% of whites on 45 items and Puerto Ricans scored within 11% of whites on 20 items. A subjective evaluation of the non-differentiating items revealed that they were more concrete or more relevant to the minority groups' experience. (For a more detailed discussion of the findings of this study, see the attached report.)

It was intended that this study would follow the subjects who were hired as Police Trainees through the Police Academy and thereby determine which test factors

BEST COPY AVAILABLE

were the best predictors of successful job performance. Unfortunately, it became impossible to conduct a formal validity study when the needs of the service forced the Police Department to defer Academy training for these new Police Trainees and to place them in varied, often non-comparable assignments.

The current study was based on the examinations for Police Trainee, Patrolman (No. 9080) and Policewoman (No. 9081) held on October 25, 1969. A substantially larger subject population was chosen to insure that we would be able to perform a validity study with this group. In pursuing this goal, we will also learn whether our predictions based on the earlier study have been confirmed. If indeed they are confirmed, a greater proportion of the items on this test should not differentiate between candidates on an ethnic basis since special efforts were made to use concrete situations in judgment items, to choose as vocabulary items words with Latin roots, and to minimize the use of academic language in the phrasing of questions and instructions.

The written test for Police Trainee, Patrolman and Policewoman was administered in two sessions (AM and PM) according to standard Department of Personnel procedures at five New York City high schools. As each candidate finished the test he was told to report to another room on the same floor. Once there, he was asked to complete a confidential personal history questionnaire. There were approximately 9,300 candidates; approximately 8,000 completed questionnaire forms which asked about their ethnic, social, economic, and educational backgrounds. This questionnaire was essentially the same in content as the one used for the earlier study, but the format was completely revised. Whereas the first questionnaire was a four-page, fill-in form, and the data had to be manually keypunched onto cards, the new form was designed to be mark-sensed by the candidate and fed through our optical scanning machine. The optical scanner would then punch the data into machine cards.

A great deal of work had to be done on these questionnaires, however, before

they could be processed by machine. Each paper was checked by an urban corps intern who removed any stray marks, cleaned up any incomplete erasures, and pointed out to the researchers any irregularities in the identifying data. In addition, about 2,000 questionnaires were coded using the socio-economic scales of the Hamburger and Shevky-Bell indices. This information will assist us in analyzing the effect of such factors as father's occupation, degree of urbanization and segregation of the candidate's neighborhood, and social rank on test ratings. After all this manual work was completed, cards were punched from the questionnaires by the optical scanner. Cards were also punched from all of the candidates' test answer papers after the second of the two lists was established in late January 1970.

The following personal data have been compiled from a preliminary analysis of the questionnaires.

Whites comprised 63% of the subject population, blacks 24%, and Puerto Ricans 11%; 7% of the subjects were born in Puerto Rico. Eighty-six percent were men and 14% were women. Most of the subjects lived in the five boroughs of New York City - 11% in Manhattan, 30% in Brooklyn, 23% in Queens, 20% in the Bronx, and 5.7% in Staten Island; only 9.6% lived in the surrounding counties of Nassau, Suffolk, and Westchester.

Twenty-six percent of the subjects were between 16 and 18 years of age; 25% were 19-21; 25% were 22-24; 16% were 25-27; and 8% were 28 or older. Three-fourths were not high school graduates (many were probably still in school at the time of the test) but 79% were employed, 80% of these full-time and 20% part-time. English was spoken in the homes of 92% of the subjects, Spanish in 5% and Italian in 3%.

We also obtained data as to where the candidates learned about this test. This information was given to Mr. DeMille. A large number of candidates (41%) found out about the test from friends and relatives who were city employees, especially policemen. The next largest group (13%) were candidates who heard about the examination through The Chief.

While making this preliminary analysis, we discovered that the optical scanner was defective and a large number of cards had been mis-punched. In some cases, we had to paint over erasures on the questionnaires with opaquing fluid because the optical scanner was overly sensitive and had registered erasures. In other cases, the scanner had in error shifted information by as much as 5 columns. We believe these errors have all been corrected. A duplicate deck has been punched to provide us with replacements in the event the original deck is damaged or broken up during analysis. We have been waiting for data processing to check these two decks. When we are certain as to their accuracy, Dr. Harold Levine of New York University will do a factor analysis comparing the biographical data with the candidates' responses to the test items.

These results should be very informative. Upon obtaining this analysis, we will measure the extent to which those items which do not appear to discriminate among candidates on the basis of their ethnic background predict performance at the Police Academy and on the job. If we can do this we will have taken an important step toward the development of tests which do not discriminate and which are also better able to predict performance.

B. Fishman
Personnel Examiner

P. Siegal
Acting Chief, Examining
Division R

THE CITY OF NEW YORK

Intradepartmental Memorandum

To: Assistant Personnel Director for Examinations

From: Phillip Siegal

Date: January 16, 1973

Subject: Status of study of Police Trainee P.D., Patrolman, No. 9080 and Policewoman, No. 9081 in connection with the [illegible] of candidates and related matter.

The written tests for Police Trainee (P.D.), Patrolman Exam. No. 9080 and Policewoman, Exam. No. 9081 were held on October 25, 1969. There were 6,826 candidates who appeared for Exam. No. 9080 and 2,499 candidates who appeared for Exam. No. 9081, making a total of 9,325 candidates. Since participation in the study was voluntary, about 7,500 candidates chose to participate in the study. Data from the confidential biographical questionnaire which these candidates completed and from their answer papers were punched on data processing cards.

There are at present 6,799 usable demographic cards and 8,528 usable test score cards. From these two data pools we were able to obtain 4,241 cases. Each case of two cards had matching identification numbers. These 4,241 sets of two cards, which include male and female candidates, constitute our working deck of cards.

The 4,241 cases have been cross-tabulated on two variables. One is the test score which has been dichotomized at 75 (the pass mark) and the other is the ethnic designation. The 4,241 cases have been compared with the 6,799 usable demographic cards and there are no significant differences with respect to demographic variables.

The population of the sample of male and female candidates was made up of 931 Black candidates, 390 Puerto Rican candidates and 2,748 White candidates.

The mean score of the Black candidates is 69.75 and the standard deviation is 16.07. The first quartile score is 62.24, the second is 72.92 and the third is 81.07

The mean score for the Puerto Rican candidates is 66.66 and the standard deviation is 16.56. The first quartile score is 57.65, the second 69.14 and the third 78.55.

The mean score for the White candidates is 78.54 and the standard deviation is 15.20. The first quartile score is 74.00, the second is 82.37 and the third quartile is 88.58.

TABLE I — 75% = passing mark

	Total of Male and Female Candidates			
	Failing	Passing	Total	% Passing
Black	485	446	931	47.9
Puerto Rican	237	153	390	39.2
White	652	2,096	2,748	76.3
Other	15	14	29	
Did not answer	80	63	143	
	1,469	2,772	4,241	

2261

Table I shows that of the total 931 Black candidates 446 (47.9%) passed; of the total of 390 Puerto Rican candidates, 153 (39.9%) passed and of a total of 2,748 White candidates, 2,096 (76.3%) passed.

TABLE II

Separated Male and Female Candidates

	Mean	Standard Deviation	Quartile 1st	2nd	3rd	N
Black Male	69.940	16.968	60.861	71.413	79.969	686
Black Female	75.189	11.539	68.125	76.864	83.350	238
Puerto Rican Male	66.008	16.776	56.375	68.727	77.861	359
Puerto Rican Female	75.519	10.261	64.875	77.250	83.625	27
White Male	78.030	15.577	73.491	81.881	88.309	2,482
White Female	83.833	10.476	80.083	85.800	90.458	258
			Did Not Indicate Sex			191
					Total	4,241

These results must be evaluated only in terms of the population which applied for the test. This was a self-selected population. This means that the three ethnic groups were not equated in terms of such factors as socio-economic background, education, interest, other employment opportunities, etc. It could very well be that if these three groups were equated on relevant factors, the statistical findings factors would have turned out to be quite different.

In this connection it can be seen in Table II that the Black, Puerto Rican and White female candidates had higher mean scores than the males of the same ethnic group. Thus, it appears that a self-selection factor is influencing the mean scores of each ethnic group. There is no evidence that women do better than men in general aptitude tests such as would be given for policemen. One strong conclusion which can be made is that those females who were better able to pass a written aptitude test applied, and the better qualified males chose not to apply. Presumably, they were seeking other employment.

The statistics for a previous Test No. 7046, Police Trainee, are as follows:

TABLE III A

	Mean	SD
White	70.5	12.9
Negro	57.9	14.6
Puerto Rican	49.1	15.7

TABLE III B

Comparison of means of Test No. 9080 (males) with Test No. 7046:

White	78.0 (9080) 70.5 (7046) 7.5 pts. rise	Blacks	67.9 (9080) 57.9 (7046) 10.0 pts. rise	P.R.	66.0 (9080) 49.1 (7046) 6.9 pts. rise

276

Table III B compares Test No. 9080 with the Police Trainee Test No. 7046. The construction of Test No. 9080 was guided by the recommendations made following a study of Test No. 7046. The data in Table III indicate that as a result of following the recommendations, there was a rise in the mean score of all candidates, with the Black candidates achieving the largest increase in mean score.

Since few candidates were appointed, we do not have a population which we can use as a criterion for validation purposes. We are therefore developing a validation method, using total and sub-test scores, individual items and demographic data. Data from Exam. No. 9080 and Exam. No. 9081 will be evaluated in terms of the recommendations which were developed following the study of Exam. No. 7046.

Phillip Siegal

Phillip Siegal, Chief
Examining Division S

276

(Statement delivered in hearings before the Municipal Police Training Council, in the week of December 10, 1973.)

I am Captain Brian F. Lavin, Director of the Equal Employment Opportunity Office of the New York City Police Department. I represent Commissioner Donald F. Cawley who requested that I express his regrets that he could not personally attend these hearings. He designated me to present the position and policy of our Department concerning the matter before this Council.

The minimum height standards which we historically required of Police Officer candidates were considered, in their time, as being consistent with the needs of our department in meeting its responsibility of safeguarding the general public.

No clear statement was ever made as to exactly why any particular height was chosen as the minimum requirement. It is probably fair to assume however, that the minimum had some relationship to the average height of males and females in the general population. A commonly held assumption was that Police Officers should not be much shorter than the average person they encounter on the street or with whom they have official police contact.

When the Law Enforcement Assistance Administration issued its height guideline on March 9th of this year we were well along in our deliberations concerning the requirements we would request the New York City Department of Personnel to establish for candidates taking the next scheduled examination for Police Officer.

Former Police Commissioner Patrick V. Murphy had created the Equal Employment Opportunity Office in January and charged it with

EXHIBIT J

insuring the department's conformity with all existing and emerging Equal Rights legislation and guidelines. One of the first items examined was the Equal Employment Opportunity Commission's, "Guidelines on Discrimination Because of National Origin". Issued in January 1972, it clearly spelled out our responsibility concerning height requirements.

The Law Enforcement Assistance Administration guideline thus only assured us that we were on the right track in closely examining for job relatedness any standard which may have discriminated against otherwise qualified minorities. We realize that minorities, as major victims of crime, have a vital stake in its prevention and control. It is our firm belief that if we do not assure the participation of all segments of the community in the agencies of criminal justice, no truly effective crime control program can be undertaken.

However, if part of the rationale behind the Law Enforcement Assistance Administration guideline is the attraction of more minorities to the police service, we suggest that such a move by itself will not necessarily accomplish that aim. No doubt it will allow some minorities and women to compete, but to seriously attempt to attract more minorities, much more than this needs be done. Exactly how much more, has been shown by our recent successful recruitment drive.

Very soon after we began our research, we discovered the fact that there is an unsettling scarcity of empirical data for either documenting or proving the validity of any height requirement.

The very few studies that have been done on this subject by others have concluded that there is no demonstrable link between the height of a law enforcement officer and his ability to satisfactorily perform the tasks required by the job. Also from the results of those limited studies it was obvious that height was not the most important factor in determining an officer's succeptability to attack.

Of singular significance in any study of the job relatedness of height requirements is that there have not been any significant number of persons in any police agency below height standards similar to our own upon which to base a comparative study of their effectiveness in the job.

We discovered also that there appears to be a clear trend in the court decisions dealing with height requirements as a selection criterion, to first require a showing that the height requirement disproportionately disqualifies members of a protected class. Once such a showing has been made, however, the burden then shifts to the employer to establish the job relatedness of the standard. This Griggs-Castro approach is the same presented by the Law Enforcement Assistance Administration guideline which thus parallels the case law in this area.

The Castro case seemed the most significant in our review. Significant in that the lower court held that the height requirement of the Boston Police Department was not unconstitutional. To quote from Judge Wyzanski's opinion, "It is probable that

while height is not determinative of fitness to be a policeman, any more than it is of fitness to be an athlete or a general, it is significantly related to fitness to be a policeman, as it is not significantly related to being a judge or a scientist".

It would be inaccurate and misleading to suggest that the Court of Appeals in the Castro Case upheld police height requirements as being reasonable. That court ruled that the record was infirm with respect to the fact that the plaintiffs had simply failed to demonstrate the disproportionate impact of the height requirement on the minorities involved. It seems clear then, that had the plaintiffs made the showing of prima facie discriminatory impact, the Court of Appeals would have required the defendant Boston Police Department to justify the minimum height requirement with convincing facts establishing a fit between the qualifications and the job. One distinct fact does emerge from the lower court decision. That being, that no statistics or evidence was cited as the basis for the District Judge's finding of "significant relationship".

This then, as we see it, is the very heart of the matter before us. Any effective challenge to height standards would, of course, require a threshold showing of disproportionate disqualification of a protected class as a trigger to close scrutiny of the requirement. If such a showing were successfully made, and we have reason to suspect that it could be, the burden would then shift to the department to justify the validity of the requirement in terms of job performance. Such justification

could only be sustained through the presentation of the results of height-job performance studies. Inasmuch as those studies have not been conducted in or by New York City in the past, there being no recognized need, we abandoned height requirements entirely for the Police Officer Exam scheduled for this coming Saturday. An Exam in which we anticipate well over 100,000 candidates to participate, and according to preliminary samplings, better than half will be members from the minority communities.

It is our sincere hope that sufficient numbers of male candidates below the former 5'7" standard will be successful in this exam. They will provide us with a population from which major, in depth studies can be undertaken to develop job performance criteria which would take into account all factors of required performance, including the total environment in which an officer is required to perform. From that research, employment standards could be established which could in no way be construed as unfairly discriminatory, and which would be compatible with the needs of effective law enforcement and those of the general public.

When the requirements for this Saturday's Police Officer Exam were announced several months ago, there were outcries in some circles that we were lowering our standards. While concerned with that type of reaction, it remains our contention that our decision to abandon height requirements, in addition to putting us in line with the guidelines and current case law, was made in

the hope that it would enable us to raise our employment standards. The abandonment of a requirement which could not meet a test of job relatedness is in no sense a lowering of standards. Ultimately, if minimum height standards are found to be non-job related, we have the option of setting such requirements as would give us a truer indication that a candidate possesses the skills and abilities necessary to be an effective Police Officer. On the other hand, the decision to abandon height requirement now was not cut in stone. If the data gathered in our research studies indicate that height is a bona fide occupational requirement, we will institute standards consonant with the results of that research. In the end, if our action has the effect of shifting reliance from non-job related selection criteria to criteria which are proven indicators of successful job performance, selection standards will have been raised, not lowered.

Thank you very much.

NOTICE OF ENTRY

Sir:- Please take notice that the within is a *(certified)* true copy of a

duly entered in the office of the clerk of the within named court on 19

Dated,

Yours, etc.,

Attorney for

Office and Post Office Address

214 East 2nd Street

Borough of Manhattan New York, N. Y. 10009

To

Attorney(s) for

NOTICE OF SETTLEMENT

Sir:—Please take notice that an order

of which the within is a true copy will be presented for settlement to the Hon.

one of the judges of the within named Court, at

on the day of 19

at M.

Dated,

Yours, etc.,

Attorney for

Office and Post Office Address

214 East 2nd Street

Borough of Manhattan New York, N. Y. 10009

To

Attorney(s) for

Index No. Year 19 74

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

ROBERT LUM,

Plaintiff

-against-

NEW YORK CITY CIVIL SERVICE COMMISSION, et. al.

Defendants.

COMPLAINT

CAROLYN M. HEFT
JOAN BERTIN LOWY

Attorney for Plaintiff

Office and Post Office Address, Telephone

214 East 2nd Street

Borough of Manhattan New York, N. Y. 10009

Tel. 777-5250

To

Attorney(s) for

Service of a copy of the within

is hereby admitted.

Dated,

Attorney(s) for

THE CITY OF NEW YORK
DEPARTMENT OF PERSONNEL
BUREAU OF EXAMINATIONS

TEST PLAN

Submit this copy before starting test preparation

Date March 12, 1970

Exam. No. 0013 Exam. Title Police Trainee (P.D.) Patrolman

No. of candidates (estimated or actual) 10,000 Type of Test Planned Objective Date of Test 5.9.70

Has a Test Plan been previously approved? [X] Yes [ ] No Date of Previous Exam. 10.25.69 Previous title, if different

Note: If this test plan is different from the previous one, indicate what the differences are and the reasons therefor.

AREAS TO BE COVERED	RELATIVE EMPHASIS
Job Situations	16 2/3%
New York City Information	16 2/3%
Vocabulary and Verbal Analogies	16 2/3%
Spatial Relations	16 2/3%
Associative Learning	16 2/3%
Arithmetic and Table Interpretation	16 2/3%
	100%

50% - new questions

50% - revised questions from Officer Series Exams. #1, #15, #16, and #17 covering the areas of Job Situations, Vocabulary and Verbal Analogies and Spatial Relations.

This test plan differs from the previous test plan as follows:

1. in line with new policy the test areas have been reduced to six major areas of equal weight;
2. Job Situations (previously referred to as Judgment) has been reduced from 24%
3. New York City Information has been increased from 8%
4. Vocabulary and Verbal Analogies have been combined and reduced from 20%
5. Spatial Relations has been increased from 12%
6. Associative Learning has been reduced from 18%
7. Arithmetic and Table Interpretation have been combined and increased from 10%
8. Reading Comprehension, previously 8%, has been deleted from the current test plan.

BEST COPY AVAILABLE

Prepared by Constance Sutherland — Approved by Unit or Division Chief M. Brownstein

Approved by or for Director of Examinations — Date

2312

THE CITY OF NEW YORK
DEPARTMENT OF PERSONNEL
BUREAU OF EXAMINATIONS

TEST PLAN

Sent it this copy [illegible] starting test prep. [illegible]

Date July 11, 1969

Exam. No. 9080 9081 Exam. Title Patrolman Policewoman

No. of Candidates (estimated or actual) 8,000 2,000 Type of Test Planned Objective Date of Test Oct. 25, 1969

Has a Test Plan been previously approved? [X] Yes [ ] No Date of Previous Exam. June 28, 1969 March 21, 1964 Previous title, if different

Note: If this test plan is different from the previous one, indicate what the differences are and the reasons therefor.

AREAS TO BE COVERED	RELATIVE EMPHASIS
Judgment	24%
Spatial Relations	12%
Verbal Analogies	8%
Vocabulary	12%
Arithmetic Reasoning	6%
Table Interpretation	4%
Reading Comprehension	8%
New York City Information	8%
Associative Learning	18%
	100%

In future please indicate condition of test is: new or used items -

At a meeting with Mr. Wiener at the end of June, attended by Mr. Baris, Mr. Siegal, Mr. Meltzer and Mr. Brownstein, Mr. Wiener directed that the written test for this examination is to consist of the above content and is to be prepared by Division [illegible] [illegible] Division K.

All new items

Bell 7/14/69

7.15.69

Prepared by Paul Heier Paul Heier

Approved by Unit or Division Chief M. Brownstein M. Brownstein

Approved by or for Director of Examinations _____ Date _____

-502

5.66

2307

THE CITY OF NEW YORK
DEPARTMENT OF PERSONNEL
BUREAU OF EXAMINATIONS

TEST PLAN

Submit this copy before starting test preparation

Date April 25, 1969

Exam. No. 9049 Exam. Title Patrolman

No. of Candidates (estimated or actual) 10,000 Type of Test Planned Written Date of Test 6.28.69

Has a Test Plan been previously approved? ☐ Yes ☐ No Date of Previous Exam. 4.5.69 Previous title, if different

Note: If this test plan is different from the previous one, indicate what the differences are and the reasons therefor.

AREAS TO BE COVERED	RELATIVE EMPHASIS
Judgment	25%
Figure Analogies and non-verbal items	10%
Verbal Analogies	12%
Civic Awareness	21%
Arithmetic	10%
Chart Interpretation	7%
Reading Comprehension	15%
	100%

Paul Heier

Prepared by Paul Heier

Approved by Unit or Division Chief M. Brownstein

Approved by or for Director of Examinations

Date

2932

5.68

THE CITY OF NEW YORK
DEPARTMENT OF PERSONNEL
BUREAU OF EXAMINATIONS

TEST PLAN

Submit this copy before starting test preparation

Date July 11, 1969

Exam. No. 9080 9081 Exam. Title Patrolman Policewoman

No. of Candidates (estimated or actual) 8,000 2,000 Type of Test Planned Objective Date of Test Oct. 25, 1969

Has a Test Plan been previously approved? [X] Yes [ ] No Date of Previous Exam. June 28, 1969 March 21, 1964 Previous title, if different

Note: If this test plan is different from the previous one, indicate what the differences are and the reasons therefor.

AREAS TO BE COVERED	RELATIVE EMPHASIS
Judgment	24%
Spatial Relations	12%
Verbal Analogies	8%
Vocabulary	12%
Arithmetic Reasoning	6%
Table Interpretation	4%
Reading Comprehension	8%
New York City Information	8%
Associative Learning	18%
	100%

In future please indicate combination of tests in use or used

At a meeting with Mr. Wiener at the end of June, attended by Mr. Baric, Mr. Siegal, Mr. Meltzer and Mr. Brownstein, Mr. Wiener directed that the written test for this examination is to consist of the above content and is to be prepared by Division [illegible]

All new items

7/16/69

7-15-69

Prepared by Paul Weier

Approved by Unit or Division Chief H. Brownstein

Approved by or for Director of Examinations _____ Date _____

CP-502 5.68

THE CITY OF NEW YORK
DEPARTMENT OF PERSONNEL
BUREAU OF EXAMINATIONS

TEST PLAN

Submit this copy before starting test preparation

Date March 31, 1970

Exam. No. 0013 Exam. Title Police Trainee, P.D., Patrolman

No. of Candidates (estimated or actual) 10,000 Type of Test Planned objective Date of Test 5.9.70

Has a Test Plan been previously approved? [X] Yes [ ] No Date of Previous Exam. 10.25.69 Previous title, if different —

Note: If this test plan is different from the previous one, indicate what the differences are and the reasons therefor.

AREAS TO BE COVERED	RELATIVE EMPHASIS
Perception	20%
General Aptitude (Verbal 20%) (Arithmetic 20%)	40%
Social Sensitivity	20%
Awareness of Urban Problems	20%
Total	100%

50% - new questions

50% - revised questions from Office Series Exams. #1, #15, and #17, and #16 covering the area of social sensitivity.

This test plan differs from the previous one as follows:

1. In line with new policy the areas have been reduced to 5 major areas of equal weight
2. social sensitivity has been added
3. social awareness of urban problems (previously referred to as New York City information has been increased from 8% and expanded to include urban problems generally)
4. vocabulary and verbal analogies have been combined and included in "General Aptitude" totaling 20%
5. arithmetic and table interpretation increased from 10%
6. reading comprehension, previously 8%, has been deleted from the current test plan
7. associative learning, formerly 16%, has been deleted from the current test plan
8. spatial relations, formerly 12%, has been deleted from the current test plan
9. "perception", 20% of this test, has been introduced into this test plan.

In actuality, the bulk of this test is manuscript is about 90% new. This is a result of the new format ordered by the Personnel Director on March 30, 1970.

Prepared by E. Meltzer

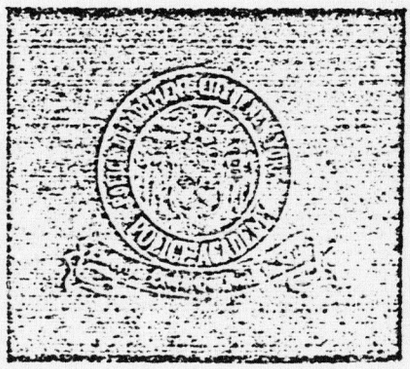
Approved by Unit or Division Chief M. Brownstein

Approved by or for Director of Examinations [signature] Date 3.31.70

DP-502 5.66

SYLLABUS

THE POLICE ACADEMY RECRUIT TRAINING SCHOOL



JULY 1967

ACADEMIC INSTRUCTION SYLLABUS
of the
RECRUITS' TRAINING SCHOOL

There are three areas of training presented to the Probationary Patrolman in the Recruit Training Program. These are physical, firearms, and academic training. The syllabuses of these three areas of recruit training are contained in this booklet.

This course of instruction is intended to present to the recruit, the fundamentals of modern police work and general knowledge necessary to make him a competent, professional police officer at the level of Patrolman.

The academic school term consists of lectures, demonstrations, playlets, practical exercises, class discussions, workshops, field trips, motion picture films, reviews, and tests. Four hours of academic instruction are presented each day that the recruits are at the Police Academy. Approximately four months of academic instruction are required to complete the training period.

The program is organized in five divisions. Each division contains several courses of related material. Each course is devoted to one phase of police work. The entire program of about three hundred hours is supplemented by approximately 100 hours of on-the-job training in field commands.

The following divisions and courses form the academic portion of the Recruit Training Program and are contained in the following pages.

DIVISION NUMBER	TITLE	COURSES	HOURS
I	Introduction to A Police Career	5	26
II	Police Procedures and Techniques	8	72
III	The Police, The Government, The Law	6	66
IV	The Police Role in Human and Race Relations	12	65
V	Demonstrations and Exercises	5	83

Issue Revised: July, 1967

POLICE ACADEMY - CITY OF NEW YORK
RECRUIT TRAINING SCHOOL

COURSE LIST

DIVISION I

"Introduction to a Police Career"

Series	Title	Hours
A	Orientation and Indoctrination	9
B	Ethics and Conduct	8
C	Equipment and Department Property	3
D	Department Organization	3
E	Reports, Records, Orders	3
		26

DIVISION II

"Police Procedures and Techniques"

A	Aided and Accident Cases	7
B	Patrol Procedures	18
C	Traffic Procedures	14
D	Investigations	3
E	Summons Procedures	8
F	Arrest Procedures	12
G	Prisoners	5
H	Police Emergencies, Disasters and Civil Defense	5
		72

DIVISION III

"The Police, The Government, The Law

A	City Government and Cooperation with Governmental Agencies	2
B	The Courts	3
C	Criminal Law and Modus Operandi	32
D	Municipal Law	12
E	Evidence	2
F	Public Morals	15
		66

COURSE LIST

DIVISION IV

"The Police Role in Human and Race Relations"

Series	Title	Hours
A	Psychology and the Police	6
B	Human Relations	5
C	Race Relations and Civil Rights	5
D	Crime and Delinquency Causation	17
E	The Constitution and Due Process	3
F	Assemblanges	3
G	Police Ethics (Chaplains)	5
H	Workshops on Human Relations	8
I	Critique Term Paper	1
J	Guest Lecturers	5
K	History of Negro in America	4
L	Puerto Rican Culture and Customs	3
		65

DIVISION V

"Demonstrations and Exercises"

A	Demonstrations	17
B	Field Trips and Duty	23
C	Practical Exercises	17
D	Review and Testing	17
E	Driver Training	9
		83

REQUIRED READINGS

1. "Crime and Race" by Marvin E. Wolfgang
   Institute of Human Relations Press
   American Jewish Committee - 1964

2. "Case Study of a Riot" by Lenora E. Berson
   Institute of Human Relations Press
   American Jewish Committee - 1966

3. "Police and the Civil Rights Act" by IACP
   International Association of
   Chiefs of Police - 1965

4. "The Puerto Ricans - Strangers, then Neighbors"
   by Clarence Senior
   Quadrangle Books - 1965

Academic:	312 hours
Physical:	192 hours
Firearms:	56 hours
Total	560 hours

DIVISION I

Orientation and Indoctrination
(26 hours)

This division is designed to indoctrinate the recruit in the mission, functions and organization of the Police Department, and with the ethics and standards with which he will be expected to comply while a member of the Department.

The material is divided into five courses and consists of twenty-six hours of instructions.

CONTENTS

Series	Hours	Title
A	9	Orientation and Indoctrination
B	8	Ethics and Conduct
C	3	Equipment and Department Property
D	3	Department Organization
E	3	Reports, Records & Orders

Division I Series A

Orientation and Indoctrination
(9 hours)

This course is intended to introduce the recruit to the Police Department and guide him in how to take notes and study. It gives him the rules of the Police Academy and the provisions of several welfare plans. During this course he is also advised of the functions of the various line organizations which he may affiliate with and requirements of producing a term paper on a subject of police interest.

1st hour — Note taking; how to study
- Taking notes in a classroom
- Keeping a notebook

Studying

2nd hour — Orientation
- Regulations of the Police Academy
- Discipline
- Payroll Information

3rd hour — Leaves and Excusals -- Part I
- How to report sick
- Medical care
- Injured member - Human bite wound
- Return to duty - Full Pay

4th hour — Leaves and Excusals - Part II
- Blood Bank
- Leaves of absence
- Vacation - Military service

5th hour — Rules and Procedures; Recruit Guide
- Instructions for amending

6th hour — Memorandum Book
- Opening memorandum book
- Sample Entry

7th hour — Review of Orientation Book

8th hour — Term Paper
- Reasons for term paper
- Reference sources
- Instructions on preparation
- Representative topics

9th hour — Personnel Records Unit
Preparation of PRU 1 (source document for personal data card)

Division I — Series B

Ethics and Conduct
(8 hours)

This course presents the code of ethics and the standards of official and personal conduct with which the recruit must comply during his period of service with the department.

1st hour — Ethics and Professionalization
- Definitions
- Ethics essential to a profession
- Code of ethics especially essential to the police profession

2nd hour — Discipline and Deportment
- Definitions
- The need for discipline
- Qualities of a well disciplined policeman

3rd hour — Official Conduct Rules - Part I
- Courtesies and salutes
- Specific rules

4th hour — Official Conduct Rules - Part II
- Continued

5th hour — General Restrictive Rules

6th hour — Suspension, Charges and Department Recognition
- Charges - suspension
- civilian complaints
- individual delinquency records
- trial procedures
- department recognition - awards

7th hour — Transfers, Details and Assignments
- Definitions
- Types
- How requested - how accomplished

8th hour — Civilian Complaint Review Board

Division I Series C

Equipment and Department Property
(3 hours)

This course deals with the rules and procedures pertaining to the procurement, use and care of personal equipment which a member of the force must possess and the care that must be exercised in the use of department buildings and equipment, including reports which must be made upon damage to such property.

1st hour Department buildings and equipment
- Department buildings
- Department equipment
- Accidents involving department property

2nd hour Uniforms and Personal Equipment
- Uniforms
- Inspection of uniforms
- Shields and cap devices

3rd hour Personal equipment
- Revolvers
- Patrol equipment
- Regulation forms
- General instructions

Division I Series D

Department Organization
(3 hours)

This course gives a brief history of the department; the administrative and functional organization and the major functions of the seperate units. The purpose of this instruction is to impart a general picture rather than to give detailed analysis.

1st hour History of Police Functions
- Anglo Saxon period
- Norman period
- Westminster period
- Colonization period
- Modern policing
- Consolidation of New York City

2nd hour Organization of the Department - Part I
- Department duties
- Order of Rank
- Functional branches

Patrol Force
Detective Division
Safety Emergency Division

3rd hour Organization of the Department - Part 2
Personnel and Administrative Services Bureau
Police Academy
Central Service Section
Central Records Office
Employees Relations Office
Office of the Chief of Staff

Division I Series E

Reports, Records and Orders
(3 hours)

An important aspect of a patrolman's work is in the realm of communication. He has to be familiar with the media through which he receives his instructions and the various types of reports he must prepare.

1st hour Reports
Explanation of official communications
When to use
Qualities of a good report
Preparation of an official report

2nd hour Forms and Records
Commonly used department forms
Instructions on forms
Certain precinct records
General instructions

3rd hour Orders
Routine department orders
General orders
Special orders
Circulars
Teletype orders
Standard operating procedures
Temporary operating procedures
Department mail

DIVISION II

Police Procedures and Techniques
(72 hours)

In this division of the recruit syllabus we present to the student the procedures to be followed in carrying out specific duties. The methods and techniques employed are explained and are coupled with practical exercises and demonstrations.

The material of this division is divided into 8 courses and consists of 72 hours of instruction. The demonstrations and practical exercises applicable to this division amount to another 12 hours (Division V), thus gicing the student a total of 84 hours of instruction in this very practical phase of police work.

CONTENTS

Series	Hours	Title
A	7	Aided and Accident Cases
B	18	Patrol Procedures
C	14	Traffic Procedures
D	3	Investigation
E	8	Summons Procedure
F	12	Arrest Procedure
G	5	Prisoners
H	5	Police Emergencies, Disters and Civil Defense

Division II Series A

Aided and Accident Cases (7 hours)

In connection with the duty of protection of life and property, the patrolman must perform certain duties in the case of persons who become ill or injured or when vehicles are involved in accidents. Cases of this nature are termed Aided and Accident Cases. To reflect the action taken by members of the force, records must be kept.

The proper preparation of these reports is dependent upon the information obtained by the patrolman assigned to these cases. This course is devoted to explanation of the need for the records and to the mechanics of preparation of the forms. Additional hours of practical exercises allows the recruit to develop the necessary skill.

1st hour — Aided Cases
- Definition
- Sick or injured persons
- Lost persons

2nd hour — Aided Cases Continued
- Dead Human Bodies
- Search of the body
- Removal from public place
- Identification of body
- Notifications
- Receipts

3rd hour — Aided Cases Continued
- Mentally ill persons
- Abandoned children
- Destitute children
- Neglected children

4th hour — Accident Cases
- Definitions
- Examinations and Investigations
- Serious injury or death
- Tow car procedures
- Animals killed or injured
- Accidents involving railroad, subway or elevated line
- Aircraft accidents

5th hour — Preparation of UF 6 card
- Preparation of UF 6
- Jurisdiction of Aided and Accident Cases

6th hour — Reports required in Aided and Accident Cases
- Number of cards required
- aided and accident record
- city liability - UF 18
- special reports required
- vehicle and traffic law reports
- reports required under traffic regulations
- transcript of accident record

7th hour — Quiz & Critique - Aided and Accident Cases

Division II Series B

Patrol Procedures
(18 hours)

The work of the patrolman is outlined in the various subjects taught the recruits during their stay at the academy. The mechanices of actual patrol, stressing the more important features, are covered in this course. The lectures are amplified by practical exercises and demonstrations and by related subjects.

1st hour — Patrol in General
- Major functions
- preparation
- proceeding to post
- general instructions
- safeguarding property
- signaling
- memorandum book
- leaving post
- notifying the station house
- post relief

2nd hour — Street Conditions - Part I
- requiring the attention of other departments
- building material in roadway
- sidewalk curb cut
- precinct survey

3rd hour — Street Conditions - Part II
- lights
- traffic hazards
- unlawful use of streets
- littering streets
- dead animals
- snow removal

4th hour — Post Conditions - Part I
- eviction proceedings
- parking meters
- derelict autos
- discarded ice boxes
- truants
- street trades
- school crossings
- children

5th hour — Post Conditions - Part II
- preservation of the peace
- unusual occurrences-diplomatic immunity
- summoning assistance
- booths
- reserve

6th & 7th hour	Motorized Patrol basic procedures operators duties automobile record recorders duties crews duties out-of-service
8th hour	Tactics Aided and Accident Cases routine procedures rules of conduct aided case at residence aided case on street vehicle accident proper lifting of aided cases
9th hour	Observation-Descriptions and Alarms observation criminal information descriptions of persons descriptions of propert. alarms road blocks code signals
10th & 11th hour	Auto Larceny classification of auto thieves modus operandi of auto thieves education of the owner recovery of stolen vehicles
12th hour	Identification Cards & Plates priviliges conveyed police cards working press card identification card emergency service card press photographers card newsreel camera vehicle card special vehicle identification card vehicle identification plates fire department plates elected officials other city departments doctors vehicles
13th hour	Responding to a Reported Crime tactics on response search of building ADT and Holmes alarm initial contact with suspect combat

14th hour	Bank Robbery points to consider plan of action observation types of bank alarms tactics of RMP crews
15th hour	Personal Safety causes of accidents accident prevention major causes of accidents developing safe attitudes
16th hour	Fugitive Felons federal procedures identification of suspects field interrogations
17th hour	Quiz and Critique - II B I to 5 incl.
18th hour	Quiz and Critique -- II B 6 to 14 incl.

Division II Series C

Traffic Procedures
(14 hours)

While handling of traffic problems is the special concern of one division of the Department, it is the duty of every member of the Force to enforce the traffic regulations and assist the Safety Division in the handling of vehicular and pedestrian traffic. This course, coupled with practical exercises and demonstrations, will equip the recruit for such duty.

1st hour	Introduction to Traffic importance of traffic control objective of traffic law enforcement state motor vehicle law enforcement techniques control techniques use of whistles
2nd hour	Department of Traffic and Safety Division functions of department of traffic organization of safety division special units care of traffic equipment
3rd hour	Traffic Regulations definitions emergency vehicles

4th hour	Traffic Regulations parking regulations parking meters off street parking facilities
5th hour	Traffic Regulations speeding right of way safety
6th hour	Traffic Regulations turns obstructing traffic park regulations - traffic
7th hour	Traffic Laws miscellaneous safety
8th hour	Traffic Regulations equipment and lights hitch hiking and soliciting play streets taxicabs peddlers omnibuses miscellaneous
9th hour	Pedestrian Regulations pedestrians subject to traffic regulations right of way of pedestrians restrictions on crossing traffic signals pedestrian control signals
10th hour	Traffic Laws registration of motor vehicles operators licenses reporting accidents leaving scene of accident compulsory vehicle inspection state highway use tax law
11th hour	Intoxicated Driving - Part I laws and presumptions
12th hour	Intoxicated Driving - Part II procedures and tests court procedures
13th hour	Quiz and Critique - II C I to 6 incl.
14th hour	Quiz and Critique - II C 7 to 12 incl.

Division II Series D

Investigations
(3 hours)

Complaints reported by the public or suspicious or dangerous conditions observed by members of the Force require investigation. This action may be taken by a detective, a plainclothesman, a superior officer or the patrolman on post. This course is designed to acquaint the recruit with the fundamentals of this phase of Police work.

1st hour	Complaints responsibility how received how recorded how referred missing persons lost property reporting a crime forms used
2nd hour	Investigations attitude toward complainant procedure at scene of crime homicide investigations investigations by a patrolman special types of investigations
3rd hour	Interrogation Techniques questioning witnesses types of witnesses interviewing victims interrogating suspects

Division II Series E

Summons Procedure
(8 hours)

A large part of the offenses against the law consists of relatively minor infractions and the persons involved are ordinary citizens who have no criminal intent. Because of this, laws have been passed and procedure established which provides for the service of a summons in lieu of an arrest.

Therefore, it is necessary that the recruit be informed of the laws which permit him to issue a summons, the violations for which he may employ this form of action and the mechanics of service and means of return. This course, together with practical exercises and demonstrations, will present this information.

1st hour	Legal Provisions and Records Summons defined NYC criminal courts act when a summons may be served summons books service of a summons
2nd hour	Traffic Summonses - Part I when returnable reporting service court appearances
3rd hour	Traffic summonses - Part II forms to be used
4th hour	Traffic Summonses, Special Reports, Records, and Traffic Summonses Control Bureau patrol precinct records special forms motor vehicle inspection regulations TSCB
5th hour	Non-Traffic Summonses court appearance forms to be used special procedures
6th hour	Service Technique non-traffic cases traffic cases pitfalls to be avoided

7th hour — Reports in lieu of summons
- U.S. govt. vehicles
- traffic violations by US mail vehicle
- diplomatic priviliges
- administrative code - civil action
- sidewalk stands

8th hour — Quiz and Critique - summons procedure

Division II — Series F

Arrest Procedure
(12 hours)

One of the most valued rights of a citizen is the right to freedom of his person. Since protection of society requires that persons lawfully charged with violating the laws of the community be given a hearing, it is necessary that the officers of the law have the power to make an arrest. Citizens also have the right to make an arrest. However, peace officers have broader powers of arrest. Both also have limitations on this power as prescribed by the law.

It is the purpose of this course to inform the recruit of his rights, duties and limitations in making arrests and when he lawfully may use force and pertinent provisions relating to the conduct of police officers.

1st hour — The Law of Arrest
- authority to arrest
- arrest defined
- who may arrest

2nd hour — Arrest under a warrant
- warrant defined
- indorsements in warrants
- disposition of defendants
- duplicate warrants

3rd hour — Arrest without a warrant
- making the arrest
- arraignment
- arrest by a private person
- use of force

4th hour — Penal Law Provisions
- crimes by public officers
- crimes against public officers

5th hour — Lawful Use of Force - Justification
- by a public officer
- by a civilian

defense persons and property
use of deadly physical force
use of force-arrests and escapes
use of force - generally
limitations - use of force
person assisting peace officer

6th hour Special Cases of Arrest - Part I
arrest for misdemeanor not committed in presence of officers;
arrests of children
material witnesses
surrender by surety
violation of parole

7th hour Special Cases of Arrest - Part II
fugitive from justice
petty violations by minors
certificate of warrant
violation of selective service act
arrests by civilian
arrests for infractions and offenses

8th hour General Provisions
diplomatic immunity
US mail employee
venue and jurisdiction
uniform close pursuit act
statute of limitations
reluctant complaint

9th & 10th hour Lawful Search and Seizure
incidentals to lawful arrest
with consent of party
with a search warrant
sidelights re: search and seizure
situations and answers

11th hour Quiz and Critique - Law of arrest
12th hour Quiz and Critique - Justitification-Use of Force

Division II Series G

Prisoners and Court Procedure
(5 hours)

Since police officers have the power and the duty to make arrests, it follows logically that there must be some established procedure for handling the prisoners from the time of arrest until disposition of the case in court.

In this course we present this procedure with emphasis on protection of society, safety of the officer and humane treatment for the prisoner.

1st hour	General Provisions frisking humane treatment search and recording transportation property of prisoner detention in court laws re: escape
2nd hour	Special Procedures children as prisoners sick or injured interview of prisoner prisoner wards material witnesses photographing & fingerprinting Sec. 552 C.C.P.
3rd hour	Court Attendance upon receipt of subpoena excusals complaint absent punctuality court complaint member of the force as a witness
4th hour	Court Attendance - Part II preparation for court appearance confer with assistant district attorney testimony use of memoranda hearsay evidence prior convictions opinion evidence physical evidence cross-examination at termination of testimony

Division II Series H

Police Emergencies and Civilian Defense
(5 hours)

The term "unusual occurrence" denotes events of serious impact upon a community. It is elementary that a far-sighted police department has established procedures for dealing with eventualities. This course presents the procedures which have been adopted by this department to cope with emergency conditions; what other units can be called upon to assist; and how to mobilize quickly and effectively to deal with the situation.

1st hour	Unusual occurrences emergency defined action to be taken temporary headquarters emergency service squad person treatening to jump explosives and suspected explosives
2nd hour	Fires types of fire alarm boxes responsibilities of members of the Force maintenance of fire lines fire report
3rd hour	Emergencies and Unusual Disorders mobilization of the force coordinated action at emergencies unusual disorder plan special zones isolation of area instructions for patrolmen
4th hour	Civilian Defense emergency duty charts air raid signals conelrad specific instructions mobilization response reports
5th hour	Mobilization at Emergency Situations plans A, B and C duties in connection therewith rapid mobilization plans

DIVISION III

The Police, The Government, The Law
(66 hours)

This division is developed so that the recruit will be acquainted with the development of legal process in society.

Included in this division will be found a discussion of the civil and criminal courts of the City, State and Federal government in which a policeman is likely to be required to appear.

The largest number of hours in the division are devoted to presenting the elements of criminal and municipal laws with which the patrolman should be familiar.

Evidence is also discussed. Description of the various types of evidence is included as well as specific steps taken to identify, handle, mark and safeguard.

CONTENTS

Series	Hours	Title
A	2	City Government and Cooperation with Governmental Agencies
B	3	The Courts
C	32	Criminal Law and Modus Operandi
D	12	Municipal Law
E	2	Evidence
F	15	Public Morals

DIVISION III Series A

City Government and Cooperation with Governmental Agencies
(2 hours)

A brief outline of the structure of the municipal government and a survey of the means of cooperation with the various governmental agencies are presented in this course.

Policemen should be familiar with the agency which employs them - in this case, the City of New York. They should also be familiar with the extent of cooperation extended to other New York City departments as well as New York State and Federal agencies.

1st hour — Governmental Structure of New York City
- the mayor
- the city council
- the board of estimate
- principal departments of city government
- how laws are enacted
- the expense budget

2nd hour — Cooperation with Federal, State and City Agencies
- cooperation with federal agencies
- cooperation with New York State agancies
- cooperation with other N.Y.C. agencies

DIVISION III Series B

The Courts

When men violate the laws of society, they are brought before judicial authorities for trial and upon conviction, sentence is imposed. The purpose of this course is to outline the court system and the jurisdiction of the various courts.

1st hour — Introduction to the Courts
- definition of legal terms
- arraignment
- order of trial
- contempt of court

2nd hour — Criminal Court of the City of New York
- organization and jurisdiction

3rd hour — Federal and State Courts
- grand jury
- petit jury
- N.Y.S. Supreme Court - criminal and civil
- appellate division of Supreme Court
- court of appeals
- federal courts

DIVISION III Series C

Criminal Law and Modus Operandi
(32 hours)

The elements of the Criminal Law of the State of New York are explained in this course so that the recruit will have at his disposal the means of recognition of the elements of these crimes and thus aid in the apprehension of violators of these laws. This information is also necessary for a policeman in preventing crime. The presentation emphasizes mainly those sections of the Penal Law with which the policeman will have most frequent contact.

1st hour	Introduction to Criminal Law historical background social basis of criminal law features of the Penal Law
2nd hour	Penal Law Definitions offenses culpability attempt to commit a crime lack of criminal responsibility
3rd hour	Assault and Related Offenses definition of terms assault - 3 degrees menacing reckless endangerment promoting suicide attempt
4th hour	Disorderly Conduct and Harassment disorderly conduct harassment aggravated harassment public intoxication falsely reporting an incident
5th hour	Loitering, Unlawful Assembly, Riot loitering - violation unlawful assembly riot - felony riot - misdemeanor inciting to riot
6th hour	Burglary and Related Offenses definition of terms criminal trespass - 3 degrees burglary - 3 degrees possession of burglars' tools

7th hour	Larceny definition of terms unlawful taking - means larceny - 3 degrees determing value
8th hour	Other Offenses Relating to Theft unauthorized use of vehicle theft of services jostling fraudulent accosting fortune telling fradulently obtaining signature
9th hour	Robbery and Criminal Possession definition of terms robbery - 3 degrees criminal possession stolen property felony & misdemeanor presumptions
10th hour	Dangerous Weapons - Part 1 definition of terms possession weapons & dangerous instruments no intent to use unlawfully with intent to use unlawfully by person under 16 by person - not a citizen
11th hour	Dangerous Weapons - Part 2 manufacture-transport defaced weapons presumptions of possession exemptions certain wounds to be reported rifles and shotguns - NYC Administrative Code
12th hour	Homicide, Abortion & Related Offenses - Part 1 definition of homicide definition of applicable terms murder attempted murder
13th hour	Homicide, Abortion & Related Offenses - Part 2 manslaughter - 2 degrees criminally negligent homicide abortion - 2 degrees self abortion - 2 degrees issuing abortional articles
14th hour	Arson, Criminal Mischief & Related Offenses definition of terms arson - 3 degrees criminal mischief - 3 degrees criminal tampering - 2 degrees reckless endangerment of property refusing to yield party line

15th hour — Kidnapping and Related Offenses
- definition of terms
- unlawful imprisonment - 2 degrees
- kidnapping - 2 degrees
- custodial interference
- federal law re: kidnapping
- modus operandi of kidnappers

16th hour — Sex Offenses - Part 1
- definition of terms
- rape - 3 degrees
- sexual misconduct
- lack of consent
- corroboration
- affirmative defense

17th hour — Sex Offenses - Part 2
- definition of terms
- sodomy - 3 degrees
- consensual sodomy
- sexual abuse - 3 degrees

18th hour — Bribery and Perjury
- definition of terms
- bribery and bribe receiving
- rewarding official misconduct
- giving/receiving unlawful gratuities
- perjury - definitions
- perjury - 3 degrees
- compounding a crime
- compromising a crime

19th hour — Anticipatory Offenses
- criminal solicitation - 3 degrees
- criminal facilitation - 3 degrees
- conspiracy - 4 degrees
- criminal liability for conduct of another

20th hour — Dangerous Drug Offenses
- definitions
- identification of drugs-narcotics
- criminal possession of drugs - 4 degrees
- presumptions
- criminal selling of drugs
- possession of hypodermic instrument

21st hour — Election and Sabbath Laws
- applicable provisions election law
- general duties of patrolman
- sabbath laws (general business law)
  - defined
  - public traffic on Sunday
  - public sports and exercises
  - local option in certain cities

22nd hour	Quiz and Critique: Assault & Related Offenses
23rd hour	Quiz and Critique: Offenses Against Public Order
24th hour	Quiz and Critique: Burglary & Related Offenses
25th hour	Quiz and Critique: Offenses Relating to Theft
26th hour	Quiz and Critique: Dangerous Weapons
27th hour	Quiz and Critique: Homicide and Abortion
28th hour:	Quiz and Critique: Arson, Crminal Mischief, Criminal Tampering, Kidnapping
29th hour	Quiz and Critique: Sex Offenses
30th hour	Quiz and Critique: Bribery and Perjury
31st hour	Quiz and Critique: Anticipatory Offenses
32nd hour	Quiz and Critique: Election and Sabbath Laws

DIVISION III Series D

Municipal Laws
(12 hours)

Much of the time of a patrolman is spent on the enforcement of local laws commonly referred to as ordinances. The purpose of this course is to acquaint the recruit with the provisions of these laws and the intelligent enforcement thereof in his daily work. These laws are found in the Administrative Code, the Traffic Regulations and regulations of the various city departments.

1st hour — Amusements and Exhibitions
- exhibitions and performances
- motion pictures
- common shows
- children in theatres
- dance halls, catering establishments
- cabarets, coffee houses

2nd hour — Firearms, Air Guns, Explosives & Fireworks
- explosives
- fireworks
- firearms
- gunsmiths
- dealers in firearms
- air guns, toy guns and hunting arrows

3rd hour	Department of Licenses general provisions billiard and pool table bowling alleys junk dealers second hand dealers peddlers laundries
4th hour	Division of Licenses licenses issued by Police Department hack drivers bureau hack owners bureau pistol license bureau tow cars sound devices noise
5th hour	Health Code animals reportable diseases drugs and medicines food and drink sidewalks and streets births and deaths miscellaneous provisions
6th hour	Streets and Street Conditions general provisions advertising encumbrances sidewalks snow and ice on sidewalk smoke and abatement unlawful construction
7th hour	Park Regulations - Part 1 definitions scope of regulations prohibited uses disorderly conduct gambling explosives and weapons aviating camping
8th hour	Park Regulations - Part 2 regulated uses protection of animals animals at large fires games loitering alcoholic beverages

9th hour — Civil Law - Part 1
- contracts defined
- essentials of contracts
- legal tender
- contract disputes
- remedies

10th hour — Civil Law - Part 2
- city marshalls
- powers
- limitations
- authority to act
- property executions
- warrants of attachment
- confessions of judgment
- role of the police

11th hour — Quiz and Critique: III D 1 to 5 inclusive

12th hour — Quiz and Critique: III D 6 to 10 inclusive

DIVISION III — Series E

Evidence
(2 hours)

The policeman should be generally familiar with the rules of evidence as observed in criminal courts. He should know what constitutes evidence, how to identify evidence, how to mark, handle and safeguard evidence. These matters are dealt with in this course.

1st hour — Introduction to Evidence
- definition
- precautions
- safeguarding and handling
- property clerk
- specific types of evidence
  - threatening letters
  - counterfeit money
  - weapons and ammunition
  - narcotics
  - alcoholic beverages
  - poisoned food

2nd hour — Evidence in Homicide Cases
- medical examiner
- clothing on deceased
- identification of deceased
- dying declarations
- police laboratory
- laws pertaining to evidence

Public Morals
(15 hours)

Violations relating to prostitution, gambling and alcoholic beverages are considered to be offenses against public morals. These violations pose special social and enforcement problems. Their enforcement is specifically delegated to members of the force not wearing uniforms. However, all members of the force must take action for violations as this division of instruction will indicate.

1st hour — General Provisions
- scope and extent
- common fallacies
- why enforcement is difficult
- departmental policy
- responsibilities - members of force
- suspected premises
- steps taken to improve enforcement
- new laws
- when gambling is legal; when illegal

2nd hour — Reports, Records and Procedures
- suspected place report
- known gambler file
- liability notices
- raided premises
- instructions for raided premises

3rd hour — Gambling Offenses - Part 1
- definition of terms
- bookmaking
- lottery and policy
- types of policy
- promoting gambling - 2 degrees
- possessing gambling records - 2 degrees

4th hour — Gambling Offenses - Part 2
- bingo games
- slot machines - definition of terms
- possession of gambling devise
- types of slot machines
- unlawful dice and card games

5th hour — Prostitutional Offenses
- prostitution
- patronizing a prostitute
- promoting prostitution - 3 degrees
- corroboration
- federal Mann Act
- types of prostitutes
- permitting prostitution

6th hour	Liquor Laws definitions prohibited sales prohibited hours of sales regulations for licensed premises restrictions on licensed premises transportations special permits
7th hour	Obscenity & Related Offenses definition of terms obscenity - misdemenaor obscenity - presumptions disseminating indecent material to minors public lewdness
8th ,9th and 10th hours	Laws and Modus Operandi of Bookmaking Bookmaking - felony Bookmaking - misdemeanor Modus Operandi commission booking services wirerooms handbooks horserooms organization
11th,12th and 13th hours	Laws and Modus Operandi of Policy Policy - Felony Policy - Misdemeanor Modus Operandi lottery sales odds and pay-off determing the number participants written records organization - syndicate
14th hour	Quiz and Critique: Gambling Offenses
15th hour	Quiz and Critique: Prostitutional Offenses, Liquor Laws and Obscenity Offenses.

DIVISION IV

The Police Role in Human and Race Relations
(65 hours)

It is essential that police recruit training include subject matter which will provide a better understanding of human behavior and which will develop proper attitudes on the part of police and which are consistent with the human relations concepts of professional police performance. Towards that end, this division is specifically oriented towards developing a professional police officer who can adopt proper attitudes toward himself and the public whom he serves.

In order to increase the scope of understanding of police-human relations concepts, various instructional techniques are utilized. Through the media of lectures, discussions, films, workshops, required readings and research projects, the recruit is exposed to those ideas, opinions, points of view and conclusions which are basic to the formulation and adoption of desirable attitudes.

CONTENTS

Series	Hours	Title
A	6	Psychology and the Police
B	5	Human Relations
C	5	Race Relations and Civil Rights
D	17	Crime and Delinquency Causation
E	3	The Constitution and Due Process
F	3	Assemblages
G	5	Police Ethics
H	8	Workshops on Human Relations
I	1	Critique of Term Paper
J	5	Guest Lecturers
K	4	History of Negro in America
L	3	Puerto Rican Culture and Customs

DIVISION IV Series A

Psychology and the Police
(6 hours)

The effectiveness of modern police work is dependent upon the extent of cooperation between the police and the public. The manner in which we reason, the emotions which influence our actions and the responses which we make to various stimuli all serve to control the quality of police service. This course is designed to present the elements of these factors to the recruit.

1st hour	Personal Element in Police Work social factors entering into police-public relations. police attitudes
2nd hour	Groups of People tensions and minority group behavior crowds
3rd hour	Minorities in our Society discussion of the social status, contributions and goals of minorities
4th hour	Handling Disturbed Persons department procedures hospital rules attitude of members of force use of baton violent patients attempted suicides
5th and 6th hours	Film - " Booked for Safekeeping" showing of film discussion of procedures for handling the disturbed person

DIVISION IV Series B

Human Relations

Cooperation is bred by understanding. Policemen who understand how people react to police actions and to the activities and conditions of daily life, are better fitted to cope with the problems facing them in serving the public.

This course is intended to give the recruit a better understanding of how to get along with people, how to obtain willing cooperation, and how to serve the City of New York more effectively.

Human Relations (con'd)

1st hour — The Public
- the law abiding citizen
- minor violations - causes and corrective measures.
- the criminal element

2nd hour — The Police
- department attempts to provide good police service.
- patrolman's place in the program
- traints of the competent policeman

3rd hour — Courtesy
- obtaining public approval
- possible pit-falls
- the effect of courtesy
- telephone courtesy

4th hour — The Press
- function of the press
- freedom of the press
- department policy
- limitations on releasing information

5th hour — Police Department's Program of Public Relations
- conducting investigations
- public opinion
- means at hand
- bureau of public information

DIVISION IV — Series C

Race Relations and Civil Rights (5 hours)

Subscribing to the doctrine that there are no second class citizens, the policeman is faced with the problems of overcoming prejudices which he encounters in his association with the public. It is essential that the policeman be keenly aware of the sociological problems existing in the community; that he appreciate that we as citizens, share a common heritage regardless of race, color, creed or national origin.

In this series of instruction, these matters are stressed.

1st hour — Racial Prejudice and Common Sense
- the nature of prejudice
- origin of harmful prejudice
- basis for prejudice

2nd hour	Racial Prejudice & Federal Laws civil rights U.S. constitution federal civil rights activity
3rd hour	Federal Civil Rights Law of 1964 an examination of the federal civil rights law and its implications for police at the local level.
4th hour	State Anti-Discrimination Laws Civil Rights Laws of New York State New York State Constitution Penal Law provisions Labor, Civil Service, Military, Correction, Public Service Laws enforcement and prevention
5th hour	Quiz and Critique: Series A,B and C - Division IV

DIVISION IV Series D

Crime and Delinquency Causation
(17 hours)

The place of the child in community life has become an increasingly more important subject in police work. The factors leading to delinquency, the social deprivation of ghetto life and its effects upon the delinquent and delinquency, the units of the department charged with dealing with children and the laws pertaining to children and to adults in their relationship with children are all discussed in this series.

1st hour	Delinquency Causation - Part 1 social factors family problem areas and the ghetto modern culture
2nd hour	Delinquency Causation - Part 2 biological factors psychological factors description of a delinquent
3rd hour	Child Development and Psychology definition of personality formation of personality stages of development adolescent culture delinquent and non-delinquent

4th hour	Introduction to Children statistics on delinquency juvenile delinquent person in need of supervision runaways neglected children
5th hour	Laws and Procedures Re: Children - Part 1 review of laws and defintions action taken re: delinquent children duties of officer - taking child into custody abandoned, destitute, neglected child child - subject of a crime
6th hour	Laws and Procedures Re: Children - Part 2 wayward minor youthful offenders related problems of minors under 21
7th hour	Laws and Procedures Re: Children - Part 3 endangering welfare of a child abandonment of a child non-support of a child sexual abuse re: children - 3 degrees
8th hour	Laws and Procedures Re: Children - Part 4 unlawfully dealing with child children in licensed premises skating rinks bowling alleys pawnbrokers-junk dealers marriage law law of support compulsory education law street trades and carrier boys
9th hour	Laws and Procedures Re: Children - Part 5 observing children in licensed premises special cases involving children preparation of Y.D. 1 when Y.D. 1 (juvenile report) prepared
10th hour	Delinquent Gangs organization social customs pattern of behavior
11th hour	Treating the Youth Gang police department program youth board street club program Board of Education

12th hour — Family Court
- jurisdiction
- purpose
- organization
- procedures

13th hour — Institutions for Care of Delinquency
- current methods for treatment
- institutional treatment
- temporary detention
- list of institutions for care of delinquent

14th hour — Police Department Delinquency Prevention Program
- units specifically handling youth
  - youth investigation bureau
  - youth records unit
  - youth squads
- precinct youth councils
- police athletic league

15th hour — Role of Policeman - Part 1
- prevention
- apprehension
- detection
- conduct - do's and don'ts

16th hour — Role of Policeman - Part 2
- lawful use of force
- signs of actual or potential delinquency
- tactical situations
  - ball playing
  - hydrants
  - handling the gang

17th hour — Quiz and Critique: Crime and Delinquency Causation

DIVISION IV — Series E

Constitution and Due Process (3 hours)

Since the work of police in this country must be performed within the framework of the U.S. Constitution and with due process of law, the police officer should be familiar with the provisions of this important document. This series presents the evolution of constitutional government in America, the structure of the Constitution and the main elements of due process.

1st hour — Introduction to U.S. Constitution
- England prior to 1600
- American colonial period

2nd hour — Colonies and States
- first state and national governments
- Constitutional Convention of 1787
- Constitution of U.S.

3rd hour — Bill of Rights
amendments to Constitution
Bill of Rights and due process
general interpretation
future development

DIVISION IV — Series F

Assemblages
(3 hours)

Lawful assembly is one of the constitutional guarantees of the American people, but exercise of that right produces many police problems. The multiplicity of personal contacts in such situations indicates the vital need for proper police attitudes in order to maintain a favorable police image while insuring that the rights of the public are fully preserved. The police officer must be aware of the conditions which may occur, the laws pertaining to such cases and the techniques for handling the lawful assemblages of people.

This series is designed to acquaint the recruit with the various types of lawful assemblages as they are concerned with the police function

---

1st hour — Public Assemblages in General
types of assemblages
psychology of assemblages
panic prevention and control
policing large gatherings

2nd hour — Parades and Public Meetings
parades
street meetings
street fairs

3rd hour — Strikes and Labor Disputes
laws applicable
picketing
rights of parties involved
duties of police
method of enforcement
public relations

DIVISION IV

Series G

Police Ethics
(5 hours)

This series is not designed to be religious instruction of any kind. It has proven acceptable and welcome to police officers of all faiths. It develops the moral principles underlying modern and progressive law enforcement in the United States. It also heavily emphasizes the human relations and public relations aspects of law enforcement.

The nature of these subject matters lend themselves more appropriately for presentation by members of the clergy. Therefore, this series is given by New York City Police Department Chaplains of all faiths.

1st to 5th hours	Discussions by Police Chaplains Integrity Courage Responsibility Prejudice Reputation

DIVISION IV

Series H

Workshops on Human Relations
(8 hours)

An effective vehicle for attitude improvement and change is the influence of the peer group. Workshops appropriately bring this factor into play. Through the media of informal discussion, the intent is not to give human relations knowledge per se, but rather to modify behavior patterns to those which are acceptable from the professional police-human relations points of view; specifically, to train recruits in the human relations aspects of their work.

1st and 2nd hours	"Open End" Workshop class discussion in the general area of human relations and civil rights.
3rd thru 6th hours	"Case Method" Discussions presentation of 4 real life episodes which involve specific police officer human relations attitudes, knowledge, skills and decisions. Human relations points of view given serious consideration.
7th and 8th hours	Human Relations Research Project class committees research specific areas of human relations; collate data; prepare report and present to class as panel discussants.

DIVISION IV

Series I
(1 hour)

During the early stages of academic instruction, each recruit is assigned a topic for the preparation of a term paper which must be completed before the termination of the course. Objectives are to encourage the practice of going to authoritative sources for information and to acquaint the student with diverse points of view within this important area of human relations. It will also require the recruit to logically analyze material in order to present his findings in a clearly written term paper.

1st hour	Critique Term Paper subjects for research include several concerned with human relations.

DIVISION IV

Series J
(5 hours)

Qualified guest lecturers have volunteered to participate in the academic program. Their services are utilized in conjunction with instruction which is related to their special fields.

Consequently, such instruction is enhanced by the added experience and knowledge which these lecturers can offer. At the same time, the student is exposed to the civilian point of view as it is related to various police activities.

1st to 5th hours	Guest Lecturers related material

DIVISION IV Series K

A History of the Negro in America

This series is devoted to film presentations coupled with discussion. The purpose of which is to provide accurate information about the history of the Negro in America and to show the vital part played by the Negro in the making of America. This media serves to present fresh viewpoints which stimulate new thinking and deepen insights into a centuries old problem.

1st to 4th hours

*Film and Discussion
out of slavery - 1619 to 1860
civil war and reconstruction - 1861 to 1877
freedom movement - 1877 to present

* Film: " A History of Negro in America"
Parts 1,2 and 3 - McGraw-Hill Text Films

DIVISION IV Series L

Culture and Customs of the Puerto Rican
(3 hours)

Film presentation coupled with discussion. The purpose is to show the contributions which have been made by the culture of Puerto Rico. Also focuses upon the problems of Puerto Ricans which are encountered in adjusting from a rural to an urban community as they relate to existing customs and language barriers.

1st to 3rd hours

Presentation, Film and Discussion
culture of Puerto Rico
family patterns
education
Puerto Rican in New York City
*film and discussion

* Film: " Beyond the Valley"
Dist. by: Commonwealth of Puerto Rico
Migration Division, Dept. of Labor

REQUIRED READINGS

"Crime and Race" by Marvin E. Wolfgang; Institute of Human Relations Press-American Jewish Committee - 1964

"Case Study of a Riot" by Lenora E. Berson; Institute of Human Rel. Press-American Jewish Committee - 1966

"Police and the Civil Rights Act" by International Association Chiefs of Police - 1965

" The Puerto Ricans-Strangers,then Neighbors" by Clarence Senior Quadrangle Press - 1965

DIVISION V

DEMONSTRATIONS AND EXERCISES
(82 hours)

Five objectives exist in the presentation of this division. The first of these is to augment the lecture and home study exercises by actual demonstrations of techniques and methods and by field trips to appropriate courts, the morgue and police headquarters. The second objective is to afford the student an opportunity to apply the knowledge imparted to him by the performance of practical exercises. The third intent of the division is to set aside time for adequate review of important principles. The fourth objective is to test the student's knowledge during, and at the end of the school term. The fifth aim is to make certain that all recruits are indoctrinated in the facets of careful driving so as to minimize personal injury and property damage.

CONTENTS

Series	Hours	Title
A	17	Demonstrations
B	23	Field Trips
C	16	Practical Exercises
D	17	Review and Testing
E	9	Driver Testing

DIVISION V — Series A

Demonstrations
(17 hours)

The amount of material presented for absorption by the students is very extensive. In order to further clarify the subject matter presented in the classroom, this course is designed, through the mechanisms of demonstrations, films, and role playing to supplement the instructions given in the classroom.

1st hour — Memorandum Book
- sample memo book entries

2nd hour — Sitting Room Activity
- general instructions
- contents of sitting room
- sitting room discipline
- two way radio

3rd hour	Patrolman's Responsibility in Station House activity prior to inspection by Sgt. sergeant's inspections roll call inspection by captain posting the platoon
4th hour	Film: "Moot Court - On the Witness Stand" discussion
5th hour and 6th hour	Film: "Goof Balls and Tee" Film: " Hard Stuff" discussion of films re: narcotics
7th hour, 8th hour and 9th hour	Role Playing - Testifying in Court preparation of court affidavit mock hearing in court presentation of evidence
10th hour 11th hour and 12th hour	Preparation for Field Safety Duty Film: "Signals & Gestures in Directing Traffic" Film: " Candid Camera Shots" Film: " Officer-Violator Relationship" principles of traffic control traffic situations - problem solving demonstration and practical work in traffic control.
13th and 14th hours	Film:"Collection and Preservation of Evidence" discussion
15th 16th and 17th hours	Preparation for Field Patrol Duty foot and RMP tactics slides - discussion of slides situations

DIVISION V Series B

Field Trips and Duty
(23 hours)

Three field trips are included in the school term. The first of these is a visit to Police Headquarters; the second consists of one half day spent in a part of the New York City Criminal Court; and the last visit is to the quarters of an Emergency Service Squad.

The purpose of these visits is to make the recruit better acquainted with these places so that he has a better understanding of how they operate.

Field Trips and Duty (con'd)

1st hour 2nd hour and 3rd hour	Visit: Police Headquarters and Police Academy the photographic unit the communications bureau the bureau of criminal identification the ballistics squad the information unit the police laboratory
4th hour 5th hour and 6th hour	Visit: Criminal Court of City of New York detention pen complaint room docket clerk the bridge the courtroom the arraignment address by a judge
7th hour 8th hour and 9th hour	Visit: Emergency Service Squad Quarters mobile units of ESD explanation of equipment demonstration of apparatus visit to morgue
10th to 16th hours	Field Safety Duty in a Safety Unit supervised by P.A. instructor
17th to 23rd hours	Field Patrol Duty in a Patrol Precinct assigned to RMP car with patrol precinct patrolmen

DIVISION V

Series C

Practical Exercises

In this course, the student makes practical application of the procedures that he has learned in the classroom, with the aim in mind to make the recruit aware of the difficulties to be encountered, and to give him confidence in his ability to perform the tasks.

1st hour	An Aided Case mimeographed story of a typical aided case is given the recruit and he is required to prepare all the necessary forms. classroom critique
2nd hour	An Accident Case Practical application, in a manner similiar to previous aided case in which student is given the facts in a hypothetical accident case. classroom critique

3rd hour	A Second Aided Case (D.O.A.)
4th hour	A Second Accident Case (D.O.A.)
5th hour	A Fire Report student is required to prepare several reports concerning fires, one of which will be in formal written form.
6th hour	A Third Aided Case (Lost Child)
7th hour	A Third Accident Case (2 Persons Injured)
8th hour	Detailed Vehicle Accident Report - U.F. 6B
9th hour	The Tag Summons student is given the facts incident to a violation of the parking regulations and is required to prepare necessary forms. classroom critique
10th hour	The Personal Service Summons student is given the facts incident to a moving traffic violation and is required to prepare the necessary forms. classroom critique
11th hour	Personal Service Summons - Health Code Violation
12th hour	Patrol Equipment stanchions fire hydrants fire boxes signal boxes barriers traffic signs
13th hour	Turning Out With Platoon recruit goes through the various phases of turning out for a tour of duty. mock station house sample roll call
14th hour	Operation of 2-Way Radio & Walkie-Talkie demonstration of 2-Way radio code signals instruction in operation actual operation by recruit
15th hour	The Property Clerk preparation of Properyy Clerk's motor vehicle invoice. classroom critique

16th hour	Arrest Case - Intoxicated Driver preparation of U.F. 127 other arrest forms simulated arrest of intoxicated driver (student) and questioned by arresting officer (student) under supervision of desk officer (instructor)
17th hour	Preparation of U.F. 61 - Complaint Form classroom preparation explanation of captions

DIVISION V Series D

Review and Testing
(17 hours)

Our standards of proficiency for recruits are set forth in Chapter 20 of the Rules and Procedures.

To determine whether our standards of academic proficiency are being maintained, it is necessary to periodically evaluate recruits through written examinations, based on the instruction received in class. There are four such examinations with provision for a review session prior to the final examination.

1st, 2nd & 3rd hours	First Examination and Critique given at the end of approximately 70 hours of instruction
4th, 5th & 6th hours	Second Examination and Critique given at the end of approximately 140 hours of academic instruction. covers the 2nd 70 hour period
7th, 8th & 9th hours	Third Examination and Critique given at the end of approximately 210 hours of academic instruction. covers the 3rd 70 hour period
10th & 11th hours	Final Review given at end of term and includes material presented during the entire term
12th, 13th & 14th hours	Final Examination given at the end of approximately 290 hours of academic instruction
15th, 16th & 17th hours	Personal Evaluation - Step # 1 short talk given to class by each recruit background, education, experiences, etc. evaluation by instructor

DIVISION V Series E

Driver Training and Testing
(9 hours)

It is imperative that each recruit be made aware of the necessity to carefully operate and maintain department vehicles. To this end he is taught these principles by lecture and use of visual aids. His reactions and vision are tested by the use of modern devices and finally he is tested under actual field conditions before he is qualified as a Department Operator.

1st, 2nd & 3rd hours	Operation and Care of Department Vehicles instruction by member of Motor Transport Unit
4th, 5th & 6th hours	Reaction and Vision Testing preparation of M.T. 13 Reaction Test Vision Testing peripheral vision driver vision telebinocular test
7th, 8th & 9th hours	Field Testing supervised by members assigned to Motor Transport Unit qualified as Operator of Department Vehicles

****

SYLLABUS
OF
RECRUIT INSTRUCTION IN FIREARMS

Probationary Patrolmen receive a minimum of 54 hours of firearms training, during which time each member fires a minimum of 340 rounds of ammunition with his .38 calibre special service revolver with a 4" barrel either Colt or Smith and Wesson. Firearms training is held at an outdoor range, with firing at the 7 - 15 - 25 and 50 yard target distances, or, during the winter months, at an indoor range at the one distance of 20 yards. The New York City Police Department silhouette and bull's eye targets are used. Probationary Policewomen, with minor exceptions, receive the same training.

Training sessions are divided into range classroom instruction in firearms related subjects, and range firing.

At all sessions classroom instruction by lecture and demonstration is given in:

1. The Revolver Manual of Arms
2. Safety regulations (General and Range)
3. Explanation of the fundamentals of Single Action or Double Action firing. Instruction is given in either Target or defensive Combat-Type firing.
4. Briefing and explanation on the type of firing to be done at the session.
5. Situational use of the firearm, including legal right and duty.
6. Questions and Answers on all phases involving firearms.
7. Dry firing.
8. A general critique after firing.

At the first session, individual records are prepared, revolvers are examined and registered, holsters examined for suitability, and further particular instruction is given in:

1. Nomenclature
2. .38 Calibre ballistics.
3. Cleaning and Care of the revolver, and this subject matter is introduced from time to time during the other sessions.

During the final session, a complete review of all prior instruction is given. Also, special instruction and demonstration is given on the cleaning and care of the revolver, followed by personal supervised cleaning of the revolver by each member.

Visual instruction includes the showing of the Department's firearms situational films and slides, and the Federal Bureau of Investigation films on Single Action and Double Action firing, and the Practical Pistol Course. Non-firing familiarization and safety instruction is given in the Department's heavy weapons and on pistols the member might encounter in the performance of his police duties.

On his first outdoor range firearms training day, the Probationary Patrolman must fire a qualifying score prior to being permitted to wear the revolver.

The next phase consists of bull's eye target shooting, single action, in slow, timed, and rapid fire. This is followed by defensive combat type training consisting of all positions as in the Practical Pistol Course using the 7 - 15 - 25 and 50 yard distances at the outdoor range; point shoulder, barricade, strong and weak hand, single and double action firing when indoors.

QUALIFICATIONS: There are five qualifications which each Probationary Patrolman must meet, before advancing to the next type of firing:

1. The original qualification to wear a revolver.

2. A single action, slow fire, bull's eye target score of 60.

3. A single action, time fire, bull's eye target score of 60.

4. As his firearms training qualification for record, a slow, time, and rapid fire, single action, bull's eye target total score of 180 out of a possible 300.

5. During the final indoor session, a combined total score of 14 hits out of a possible 20 on a silhouette target out of: 10 rounds fired double action, point shoulder, and 10 rounds fired single action, strong hand, barricade.

THE RECRUITS' TRAINING SCHOOL

OUTDOOR RANGE RECRUIT TRAINING
(6 Sessions Minimum)

SESSION	# of Rounds	TYPE OF FIRING	TARGET
1st (25 yds)	40	Slow Fire, Single Action-10 rds) Single Action for qualification to carry revolver (entry on PA-12 as per qualification). do do 10 rds) do do 10 rds) do do 10 rds) Preparation of PA 11 and PA 12. Registration of revolvers Inspection of revolvers Inspection of holsters "Safety" lecture Nomenclature Cleaning and care of revolver Ballistics Revolver Manual Single Action Lecture	Sil. Sil. Sil. Sil.
2nd (25 yds)	60	Slow Fire, Single Action-10 rds (Revised Bull's Eye do do 10 rds do do 10 rds do do 10 rds do do 10 rds do do 10 rds (Highest qualifying score recorded on PA 12) Revolver Manual "Safety" lecture Stamping Equipment F.B.I. Single Action Film Lowering of the hammer on a "Live" round Single Action review lecture	B.E. B.E. B.E. B.E. B.E. B.E.
3rd (25 yds)	60	Time Fire, Single Action-10 rds) Highest Qualifying Score Attained Recorded on PA 12. Time Fire, Single Action-10 rds) Time Fire, Single Action-10 rds) Time Fire, Single Action-10 rds) Rapid Fire, Single Action-10 rds	B.E. B.E. B.E. B.E.

THE RECRUITS' TRAINING SCHOOL

OUTDOOR RANGE RECRUIT TRAINING

SESSION	# of Rounds	TYPE OF FIRING	TARGET
3rd Cont'd (25 yds)	60	Rapid Fire, Single Action - 10 rds	B.E.
		Revolver Manual "Safety" lecture Review "Single" Action Instruct "Time" fire Instruct "Rapid" fire	
4th (25 yds)	60	Rapid Fire, Single Action - 10 rds)Highest Rapid Fire, Single Action - 10 rds)Score at- Rapid Fire, Single Action - 10 rds)tained in Slow Fire, Single Action - 10 rds)this and Time Fire, Single Action - 10 rds)last ses- Rapid Fire, Single Action - 10 rds)sion in)rapid fire)will be)recorded on)PA 12	B.E. B.E. B.E. B.E. B.E. B.E.
		Review of "Safety" Lowering hammer on a live round Duffco Machine Review Single Action-Slow-Time-Rapid	
5th (25 yds)	50	Slow Fire, Single Action - 10 rds)Test in Time Fire, Single Action - 10 rds)Single Rapid Fire, Single Action - 10 rds)Action- 180 re- quired record scored on PA 12 and total	B.E. B.E. B.E.
(15 yds)		Point Shoulder, Double Action, Strong Hand Supported - 10 rds Point Shoulder, Double Action, Strong Hand, Unsupported - 10 rds	Sil. Sil.
		Review of "Safety" Revolver Manual Review Single Action-Slow-Time-Rapid Instruct Double Action Point Shoulder Instruct Combat Load F.B.I. Film on Double Action Firing	

THE RECRUITS' TRAINING SCHOOL

OUTDOOR RANGE RECRUIT TRAINING

SESSION	# of Rounds	TYPE OF FIRING		TARGET
6th (15 yds)	70	Point Shoulder Double Action, unsupported	10 rds	Sil.
		Point Shoulder Double Action, unsupported kneeling -	10 rds	Sil.
(25 yds)		Single Action, Barricade, Strong Hand	20 rds	Sil.
		Single Action, Barricade, Weak Hand	10 rds	Sil.
(15 yds)		Point Shoulder, Double Action, supported or unsupported - (14 hits required record on PA 12) -	10 rds	Sil.
(25 yds)		Single Action, Barricade, Strong Hand (14 hits required record on PA 12) -	10 rds	Sil.
		Review of "Safety"		
		Review Double Action, Point Shoulder, supported and unsupported, and kneeling.		
		Review Combat Load		
		Instruct Strong Hand and Weak Hand, Barricade shooting.		

THE RECRUITS' TRAINING SCHOOL

INDOOR RANGE RECRUIT TRAINING

SESSION	TOTAL NUMBER OF ROUNDS	TYPE OF FIRE	TYPE OF TARGET
1* (ALL DAY)	20	10 Slow Fire	Silhouette
		10 Slow Fire	do
2nd	20	10 Slow Fire	do
		10 Slow Fire	do
3rd	20	10 Slow Fire	Bull's eye(Rev)
		10 Slow Fire	Bull's eye
4th	20	10 Slow Fire	do
		10 Slow Fire	do
5th	20	10 Slow Fire	do
		10 Slow Fire	do
6th	20	10 Time Fire	do
		10 Time Fire	do
7th	20	10 Time Fire	do
		10 Time Fire	do
8th	20	10 Rapid Fire	do
		10 Rapid Fire	do
9th	30	10 Rapid Fire	do
		10 Rapid Fire	do
		10 Rapid Fire	do
10th	20	10 Strong Hand, Point Shoulder, Double Action, Supported	Silhouette
		10 Strong Hand, Point Shoulder, Double Action, Unsupported -	Silhouette
11th	20	10 Strong Hand, Point Shoulder, Double Action, Unsupported -	Silhouette
		10 Strong Hand, Single Action, Barricade.	Silhouette
12th	20	10 Strong Hand, Single Action, Barricade.	Silhouette
		10 Strong Hand, Point Shoulder, Double Action, Unsupported -	Silhouette
13th	20	10 Weak Hand, Single Action Barricade	Silhouette
		10 Slow Fire	Bull's eye
14th	20	10 Time Fire	do
		10 Rapid Fire	do
15th*	30	10 Slow Fire) Test	
		10 Time Fire) 180 required	do
		10 Rapid Fire)	

THE RECRUITS' TRAINING SCHOOL

INDOOR RECRUIT TRAINING

SESSION	TOTAL NUMBER OF ROUNDS	TYPE OF FIRE	TYPE OF TARGET
16th(ALL DAY)	40	(10 Strong Hand, Point Shoulder,	
		(Double Action, Unsupported -	Silhouette
	- Test	(10 Strong Hand, Single Action	
	14 Hits	(Barricade	Silhouette
		10 Double Action, Rapid Fire Cadence	Silhouette
		10 Double Action, Rapid Fire Cadence	Silhouette

Minimum 340 rounds -- All .38 calibre special ammunition.

*All sessions except the 1st and 16th are half-day sessions. The 1st and 16th sessions are whole-day sessions.

*The scores of the 15th session Indoors, or the 5th session Outdoors, are used as the qualification scores for candidacy, for the Masback Trophy shoot off.

THE RECRUITS' TRAINING SCHOOL

PHYSICAL TRAINING PROGRAM

The physical training program is designed to develop a high degree of strength, endurance, agility, coordination and skill in the Police Recruit. The development and maintenance of a sound physique, supplementing the Academic and Firearms Training, will produce a highly efficient, well trained policeman, capable of performing a wide variety of duties and meeting emergency situations.

The Physical Program consists of training in such subjects as infantry drill, calisthenics, baton drill, search and frisk, boxing, unarmed defense, riot control, and first aid. Sessions are held four (4) days each week for a period of three (3) hours each. A total of one hundred ninety-two (192) hours of training is scheduled over a four (4) month period.

THE RECRUITS' TRAINING SCHOOL

PHYSICAL SCHOOL
POLICE ACADEMY

MASTER TRAINING SCHEDULE

DAY	1ST PERIOD	2ND PERIOD	3RD PERIOD
1	Orientation	Orientation	Military Drill (1)
2	Cal (1)	Baton (1)	Military Drill (2)
3	Cal (2)	Baton (2)	Military Drill (3)
4	Cal (3)	Baton (3)	Search & Frisk (1)
5	Cal (4)	Search & Frisk (2)	First Aid (1)
6	Cal (5)	Search & Frisk (3)	First Aid (2)
7	Cal (6)	Search & Frisk (Cuffs)	First Aid (3)
8	Cal (7)	Judo (1)	First Aid (4)
9	Cal (8)	Judo (2)	First Aid (5)
10	Cal (9)	Judo (3)	First Aid (6)
11	Cal (10)	Judo (4)	First Aid (7)
12	Cal (11)	Judo (5)	First Aid (8)
13	Cal (12)	Judo (6)	First Aid (9)
14	Cal (13)	Judo (7)	First Aid (10)
15	Cal (14)	Judo (8)	First Aid (11)
16	Cal (15)	Boxing (1)	First Aid (12)
17	Cal (16)	Boxing (2)	First Aid *13)
18	Cal (17)	Judo (9)	First Aid (14)
19	Cal (18)	Boxing (3)	First Aid (15)
20	Cal (19)	Judo (10)	First Aid (16) Test
21	Cal (20)	Boxing (4)	Baton (Review)
22	Cal (21)	Boxing (5)	Search&Frisk(Rev.1,2,3
23	Cal (22)	Judo (11)	Military Drill (4)
24	Cal (23)	Judo (12)	Boxing (6)
25	Cal (24)	Judo (13)	Boxing (7)
26	Cal (25)	Judo (14)	Boxing (8)
27	Cal (26)	Judo (15)	Boxing (9)
28	Cal (27)	Judo (16)	Military Drill (5)
29	Cal (28)	Judo (17)	Boxing (6)
30	Cal (29)	Riot Control (1)	Boxing (7)
31	Cal (30)	Boxing (11)	Boxing (8)
32	Cal (31)	Judo (18)	Boxing (9)
33	Cal (32)	Boxing (12)	Military Drill (5)
34	Cal (33)	Judo (19)	Riot Control (2)
35	Cal (34)	Boxing (14)	Riot Control (3)
36	Cal (35)	Judo (20)	Riot Control (4)
37	Cal (36)	Judo (21)	Boxing (13)
38	Cal (37)	Judo (22)	Military Drill (7)
39	Cal (38)	Judo (23)	Boxing (15)
40	(39)	Judo (24)	Boxing (16)
			Military Drill (8)
			Boxing (17)
			Search&Frisk(Rev.Cuffs)

THE RECRUITS' TRAINING SCHOOL

DAY	1ST PERIOD	2ND PERIOD	3RD PERIOD
41	Cal (40)	Judo (25)	Boxing (18)
42	Cal (41)	Judo (26)	Military Drill (9)
43	Cal (42)	Judo (27)	Riot Control(Review)
44	Cal (43)	Judo (28)	Boxing (19)
45	Cal (44)	Judo (29)	Boxing (20)
46	Cal (45)	Judo (30)	Boxing (21)
47	Cal (46)	Judo (31)	Boxing (22)
48	Cal (47)	Judo (32)	Military Drill (10)
49	Cal (48)	Judo (33)	Boxing (23)
50	Cal (49)	Judo (34)	Boxing (24)
51	Cal (50)	Judo (35)	Military Drill (11)
52	Cal (51)	Judo (36)	Boxing (25)
53	Cal (52)	Judo (37)	Boxing (26)
54	Cal (53)	Judo (38)	Boxing (27)
55	Cal (54)	Judo (39)	Boxing (28)
56		PHYSICAL TEST	
57		PHYSICAL TEST	
58	Cal (55)	Judo (40)	Boxing (Review 1-3)
59	Cal (56)	Judo (41)	Boxing (Review 4-7)
60	Cal (57)	Judo (42)	Boxing (Review 8-11)
61	Cal (58)	Judo (43)	Boxing (Review 12-16)
62	Cal (59)	Judo (44)	Boxing (Review 17-22)
63	Cal (60)	Judo (45)	Boxing (Review 23-28)
64	Cal (61)	Judo (46)	Baton (Review 1-2)
65	Cal (62)	Judo (47)	Baton (Review 3-4)
66	Cal (63)	Judo (48)	Military Drill (12)
67	Cal (64)	Judo (General Review)	Search & Frisk(Rev.1-2)
68	Cal (65)	Search & Frisk Review 3-4)	Military Drill (General Review)

THE RECRUITS' TRAINING SCHOOL

INFANTRY DRILL
(12 LESSONS)

Lesson 1 SCHOOL OF SOLDIER

POSITION OF ATTENTION
POSITIONS OF REST (PARADE, AT EASE & REST)
SALUTES
INTERVAL & DISTANCE (NORMAL & CLOSE)
DRESSES (RIGHT & LEFT)
INSPECTION (OPEN & CLOSE RANKS)
FACING MOVEMENTS

Lesson 2 STEPS & MARCHING

LEFT & RIGHT STEP, MARK TIME, FORWARD MARCH, HALF STEP
BACKWARD MARCH
CHANGE OF DIRECTION:
BY THE LEFT & RIGHT GUIDE, BY THE LEFT & RIGHT FLANK
TO THE REAR, GUIDES LEFT & CENTER

Lesson 3 SCHOOL OF SQUAD & PLATOON

FORM SINGLE FILE
FORM COLUMN OF TWO'S
COUNT OFF
LEFT & RIGHT OBLIQUE
COLUMN MOVEMENTS

Lesson 4 MASS COMMANDS & MILITARY COURTESY

REVIEW LESSONS 1 - 3

Lesson 5 DRILLS & CEREMONIES

Lesson 6 PEP & POSTURE DRILL

Lesson 7 PARADE FORMATIONS

SELECT COLOR GUARD

THE RECRUITS' TRAINING SCHOOL

Lesson 8 COLUMN MOVEMENTS & TURNS

Lesson 9 PASS IN REVIEW

EYES RIGHT & LEFT

Lesson 10 FORM IN LINE OF COMPANIES

IN MASS FORMATIONS FOR CEREMONIES

Lesson 11 COLUMN OF PLATOONS IN LINE

Lesson 12 FORMATION FOR CEREMONIES - GRADUATION FORMATIONS

A: SALUTES & COURTESIES

1. PARADE REST FOR CHAPLAIN SERVICES
2. CLASS PRESENTATION (SALUTE)
3. SALUTE TO NAT'L. ANTHEM & NATIONAL COLOR

NOTE: LESSONS USED IN INFANTRY DRILL, TAKEN FROM U.S. ARMY FIELD MANUAL F.M. 22.5.

THE RECRUITS' TRAINING SCHOOL

PHYSICAL SCHOOL
POLICE ACADEMY

RECRUIT CALISTHENIC PROGRAM
(64 Classes of 30 Min.)
32 Hours "

CALISTHENICS BASED ON F.M. 21-20 U.S. ARMY

1. STANDING POSITION

A. SIDE STRADDLE HOP

1st Week	15 Repetitions		
2nd "	20	"	
3rd "	25	"	
4th "	20	"	(2 sets)
5th " (To Finish)	25	"	" "

B. ALL OTHERS ----10 REPETITIONS

(1) Arm Exercises
(2) Leg Exercises
(3) Trunk Bends
(4) Trunk Twists, etc.

2. FLOOR POSITION

A. SIT-UPS (Minimum for Test --- 42 Repetitions)

1st Week	15 Repetitions		
2nd "	20	"	
3rd "	25	"	
4th "	30	"	
5th "	40	"	(2 Sets of 20)
6th " (To Finish)	50	"	(2 Sets of 25)

B. PUSH-UPS (Minimum for Test --- 22 Repetitions)

1st Week	10 Repetitions		
2nd "	12	"	
3rd "	14	"	
4th "	16	"	
5th "	18	"	
6th "	20	"	
7th " (To Finish)	30	"	(2 Sets of 15)

C. SQUAT THRUSTS (Minimum for Test --- 21 Repetitions)

1st Week		5	Repetitions
2nd "		7	"
3rd "		9	"
4th "		10	"
5th "		12	"
6th "	(To Finish)	20	"(2 Sets of 10)

3. RUNNING (Minimum for Test --- 1/2 Mile in 3 Min. & 15 Seconds)

1st Week		1	Minute
2nd "		1-1/2	"
3rd "		2	"
4th "		2-1/2	"
5th "		3	"
6th "		3-1/2	"
7th "	(To Finish)	4	"

THE RECRUITS' TRAINING SCHOOL

BATON
(6 Lessons)

Lesson 1

A) PROPER METHOD OF INTRODUCING THONG
B) PROPER GRIP
C) POSITION OF CARRY
D) PARADE REST
E) PRESENT ARMS (BATON)
F) MARCHING WITH BATON
G) GROUND BATON

Lesson 2

A) ALERT POSITION
B) ON GUARD POSITION
C) BLOWS WITH BATON
D) THRUST (TWO COUNTS)
E) THRUST (THREE COUNTS)
F) THRUST (FOUR COUNTS)
G) STRAIGHT ARM AND THRUST

Lesson 3

A) RIGHT HAND SWEEP
B) LEFT HAND SWEEP
C) BLOCKING KICK TO GROIN
D) STRANGLE HOLDS

Lesson 4

A) COME-ALONG, WITH BATON
B) DEFENSE, RIGHT HAND BLOWS TO HEAD
C) DEFENSE, LEFT HAND BLOWS TO HEAD
D) HANDLING CROWDS
E) STICK WRESTLING, RECOVERING BATON FROM OPPONENT'S GRASP

Lesson 5 REVIEW LESSONS #1 AND #2

Lesson 6 REVIEW LESSONS #3 AND #4

THE RECRUITS' TRAINING SCHOOL

SEARCH, FRISK & HANDCUFFS
(6 Lessons)

Lesson 1 TECHNIQUE OF FRISKING A PRISONER

A) INTRODUCTION
B) THE APPROACH
C) THE ACTUAL FRISK

Lesson 2 SECURING THE PRISONER BEFORE OR AFTER THE FRISK

A) DEPARTMENT REGULATIONS GOVERNING USE OF HANDCUFFS
B) MANNER OF HANDCUFFING
C) COMMON MISTAKES IN THE USE OF HANDCUFFS

Lesson 3 MULTIPLE FRISK BY ONE OFFICER AND SUMMARY

Lesson 4 TECHNIQUE OF OVERTAKING AND SEARCH OF SUSPECTED AUTO

A) INTRODUCTION
B) PROCEDURE
C) SUMMARY

Lesson 4a FOOT PATROLMAN'S APPROACH TO SUSPICIOUS AUTOMOBILE

Lesson 5 REVIEW OF FRISK AND OPPORTUNITY FOR RECRUITS TO FRISK AND APPLY HANDCUFFS TO EACH OTHER.

Lesson 6 SAME

THE RECRUITS' TRAINING SCHOOL

BOXING
(28 Lessons)

Lesson 1	Stance Footwork Straight Left to chin
Lesson 2	Review Lesson No. 1 Defenses for straight left to chin
Lesson 3	Review straight Left to chin and defenses Additional defenses for straight Left to chin
Lesson 4	Review straight Left to chin and defenses Straight Left to body Defense for straight Left to body
Lesson 5	Review straight Left to chin, body and defenses Combination blows with straight Left Actual boxing using only straight Left hand blows
Lesson 6	Review straight Left defenses and combinations Left hook to chin Defenses for Left hook to chin Actual boxing, using only straight Left hand blows
Lesson 7	Review straight Left, Left hook to chin and defenses Left hook to body Defenses for Left hook to body Actual boxing, using all Left hand blows
Lesson 8	Review Left hand blows and defenses Combinations with Left hand blows Actual boxing, using all Left hand blows & combinations
Lesson 9	Review Left hand blows, defenses and combinations Counters for straight Left, hook and uppercut Actual boxing, using all Left hand blows, counters, and combinations
Lesson 10	Review Left hand blows, defenses, counters and combinations
Lesson 11	Review Left hand blows and defenses Straight Right to chin Defenses for straight Right to chin Actual boxing, using all Left hand blows taught

Lesson 12	Review straight Right to chin and defenses Straight Right to body Defenses for straight Right to body Actual boxing, using all blows taught
Lesson 13	Review straight Right to chin, body and defenses Combination blows with Left and straight Right hand punches Actual boxing, using all blows taught
Lesson 14	Review straight Right hand blows and defenses Counter blows against straight Right blows Actual boxing, using all blows taught
Lesson 15	Review straight Right hand blows and defenses Right hook to chin Defenses against Right hook to chin Actual boxing, using all blows <u>except</u> Right hooks
Lesson 16	Right hook to chin and defenses Right uppercut to body Defenses for Right uppercut to body Actual boxing, using all blows taught
Lesson 17	Review Right hook, uppercut and defenses Combination blows Actual boxing using all blows
Lesson 18	Review Left and Right blows and defenses Combination blows Actual boxing, using all blows
Lesson 19	Review Left and Right hand blows and defenses Combination blows Actual boxing
Lesson 20	Review Left and Right hand blows and defenses Combination blows Actual boxing
Lesson 21	Review Left and Right hand blows and defenses Combination blows Counter blows Actual boxing
Lesson 22	Review Left and Right hand blows and defenses Combination blows Counter blows Actual boxing

THE RECRUITS' TRAINING SCHOOL

Lesson 23	Review Counter and Combination blows Actual boxing
Lesson 24	Review Counter and Combination blows Actual boxing
Lesson 25	Review blocking a straight Right to the Head Review slipping, ducking, feinting and clinching
Lesson 26	GENERAL REVIEW OF BOXING PROGRAM
Lesson 27	Hitting Drill - All blows taught Blocking Drill - All blows taught
Lesson 28	Actual boxing

NOTE: Reference used for all boxing lessons is the Army Field Manual FM 21-20

THE RECRUITS' TRAINING SCHOOL

UNARMED SELF-DEFENSE
(48 Lessons)

Lesson 1	Orientation--Student Participation (a) Vulnerable parts of the body (b) Weapons and Attacks (c) Balance and Leverage workout
Lesson 2	Forward Roll Backward Roll Side, Break fall
Lesson 3	Same as Lesson 2
Lesson 4	Leg trip, (by the #'s)
Lesson 5	Leg trip, (by the #'s)
Lesson 6	Leg trip, (by Command)
Lesson 7	Leg trip, (by Command)
Lesson 8	Body drop, (by the #'s)
Lesson 9	Body drop, (by the #'s)
Lesson 10	Body drop, (by Command)
Lesson 11	Body drop, (by Command)
Lesson 12	Leg trip and Body drop (combinations)
Lesson 13	Same ---------
Lesson 14	Same ---------
Lesson 15	Same ---------
Lesson 16	Rear Leg take down
Lesson 17	Same ---------
Lesson 18	Front tackle
Lesson 19	Same --------
Lesson 20	Front choke, side headlock
Lesson 21	Same --------

THE RECRUITS' TRAINING SCHOOL

Lesson 22	Rear choke (by the #'s) Use Right arm Only
Lesson 23	Same --------
Lesson 24	Rear choke (by Command)
Lesson 25	Rear choke (by Command) Right and Left Arms
Lesson 26	Bridging
Lesson 27	Bridging
Lesson 28	Wristlocks
Lesson 29	Same --------
Lesson 30	Same --------
Lesson 31	Restraining Holds (Hammerlock)
Lesson 32	Same --------
Lesson 33	Same --------
Lesson 34	Restraining Holds (Front Arm Bar)
Lesson 35	Restraining Holds (Rear Choke)
Lesson 36	Same --------
Lesson 37	Disarming (Gun)
Lesson 38	Same --------
Lesson 39	Disarming Knife, Club, Gun (kicking)
Lesson 40	Removing unwilling persons (a) Chair (b) Door (c) Floor
Lessons 41 to 48	REVIEW FOREGOING LESSONS

THE RECRUITS' TRAINING SCHOOL

RIOT CONTROL
(4 Lessons)

1st Hour CROWD & MOB PSYCHOLOGY - LECTURE

2nd Hour SQUAD FORMATIONS - DEMONSTRATION & PARTICIPATION

(A) Wedge
(B) Diagonal
(C) Line

3rd Hour PLATOON FORMATIONS - DEMONSTRATION & PARTICIPATION

(A) Wedge
(B) Diagonal
(C) Line

4th Hour GENERAL REVIEW - LECTURE & DEMONSTRATIONS

THE RECRUITS' TRAINING SCHOOL

FIRST AID TRAINING COURSE
(16 Lessons)

Lesson 1 INTRODUCTORY REMARKS

Define:
(a) Purpose
(b) Need
(c) Accident Scene

(1) Frequency of accidents and illnesses
(2) Need for immediate care
(3) Public relations
(4) Obligation to Public
(5) Self-satisfaction

Lesson 2 SHOCK
(a) Nature of Shock
(b) Causes
(c) Contributory factors
(d) Dangers of Shock
(e) Symptoms
(f) Pitfalls
(g) First Aid and Public Relations

Lesson 3 CONTROL OF BLEEDING
(a) Types of wounds (P A I L)
(b) Direct pressure, digital pressure, elevation tourniquet
(c) Infection
(d) Abdominal wounds
(e) Gunshot wounds
(f) Foreign object embedded in wound
(g) Dog bite
(h) Fractures

Lesson 4 ARTIFICIAL RESPIRATION (MOUTH TO MOUTH, BACK PRESSURE ARM LIFT)

(a) Reasons for artificial respiration
(b) Types of Cases:
(1) Drowning
(2) Electric shock
(3) Gas Poisoning
(4) Anesthetic (drugs)
(5) Chest Compression
(6) Hanging
(7) Choking
(8) Chest Puncture wound
(9) Prolonged exposure
(10) Excessive Alcohol

THE RECRUITS' TRAINING SCHOOL

Lesson 5 ARTIFICIAL RESPIRATION - PARTICIPATION

(a) Manual Methods

Lesson 6 POISONS

(a) Poisons used
(b) Symptoms
(c) First Aid and prevention

Lesson 7 EFFECTS OF EXCESSIVE HEAT AND UNCONSCIOUSNESS

(a) Sunstroke or heat stroke
(b) Heat Exhaustion and heat cramps
(c) Symptoms
(d) First Aid and prevention
(e) Alert for medical disks, bracelets & Tags

Lesson 8 COMMON MEDICAL EMERGENCIES

(a) Heart Failure
(1) Symptoms
(2) First Aid
(b) Apoplexy
(1) Symptoms
(2) First Aid
(c) Epileptic convulsion
(1) Symptoms
(2) First Aid
(d) Pain in the Abdomen-Appendicitis
(1) Symptoms
(2) First Aid
(e) Convulsions in Children
(1) Symptoms
(2) First Aid

Lesson 9 BURNS

(a) Classification
(b) Prevention
(c) First Aid
(d) Handling of Psycho

Lesson 10 TRANSPORTATION

(a) Importance
(b) Methods

THE RECRUITS' TRAINING SCHOOL

Lesson 11	EMERGENCY CHILD BIRTH (a) Film (b) Lecture
Lesson 12	ARTIFICIAL RESPIRATION (a) Film (b) Lecture
Lesson 13	REVIEW Lessons 1, 2, 3
Lesson 14	REVIEW Lessons 4, 5, 6
Lesson 15	REVIEW Lessons 7, 8, 9
Lesson 16	TEST

TEXT BOOK: AMERICAN NATIONAL RED CROSS - 4th EDITION